Què diem de nosaltres mateixos quan volem presentar-nos davant dels altres com un 'producte' atractiu?

Per: Josep Anton Charles Ortega
Dirigit pel Doctor: Joaquim Moré López

TFM
Màster Universitari en Ciència de Dades
U.O.C. Universitat Oberta de Catalunya.

3 gener 2021

1 Abstract.

This study analyzes the different discursive strategies, understood as a combination of themes, that individuals use when presenting themselves to others as an "attractive product", in the context of an online dating platform. The analysis is based on the free text in each individual's self-summary, from which their specific words are obtained. Using LDA (latent dirichlet allocation), the number of themes discussed and the keywords in each theme are determined. To label topics found in self-summaries, we extracted the semantic relationship between said keywords with the help of a word2vec algorithm created with the total speeches that we visualize with a TSNE. When we have determined the topics, we can decompose the discourse of an individual into a linear combination of topics used. In this decomposition, the weight of each topic are interpreted its relevance within the discourse, referred to as the discursive profile. Then using a cluster analysis of these discursive profiles, different communication patterns are determined, referred to as discursive strategies. Finally, the usage of these strategies among different individuals is analyzed, determining who uses what discursive strategies.

En aquest estudi s'analitzen les diferents "estratègies discursives", enteses com a combinació de "temes", que els individus fan servir quan es presenten davant dels altres com a "producte" atractiu, en el context d'una plataforma de cites online. Es parteix del "text lliure" que hi ha en el "discurs" de cada individu, del qual s'obté les seves "paraules". Es determina el número de "temes" sobre els que es parla i es busca les "paraules clau" de cada "tema" amb LDA (latent

dirichlet allocation). Per a etiquetar els "temes" extraiem la relació semàntica entre les "paraules clau" amb l'ajut d'un word2vec creat sobre el total de "discursos" que visualitzem amb un TSNE. Quan tenim determinats els "temes" podem descompondre el "discurs" d'un individu com una "combinació lineal" dels temes utilitzats que denominarem "perfil discursiu". Els pesos s'interpreten com la rellevància del "tema" dintre del "discurs". A continuació, amb l'ajut d'una anàlisi cluster, es determinem els diferents "patrons", que denominarem "estratègies discursives". Per últim, s'analitza quines "estratègies discursives" utilitzen les diferents "tipologies" d'individus.

2 Paraules clau.

NLP, processament del llenguatge natural, detecció de temes, cites online.

3 Motivació del projecte.

Amb aquesta investigació es vol determinar des d'un punt de vista general quins són els principals temes dels quals una persona parla quan es vol presentar davant dels altres com un "producte" atractiu.

Des d'un punt de vista psicològic sembla interessant determinar quins són aquests temes principals; quines són les diferents "estratègies discursives", enteses com a combinació de temes; i si existeixen diferències entre les "estratègies discursives" en funció dels perfils d'individus.

Interessa analitzar el "text lliure" com a font d'informació d'aspectes rellevants, en aquest cas el que utilitzen les persones per descriure's, per mostrar els seus punts forts, el que busquen, etc. És important destacar que el fet de ser "text lliure" dóna llibertat a l'usuari, i permet fer-ho com ell consideri oportú i mostrar-nos allò que ell considera rellevant.

Des del punt de vista professional, quan es vol avaluar un producte, idea, etc s'utilitzen qüestionaris, en els quals l'investigador decideix quines preguntes són rellevants o no, independentment de la rellevància que tinguin per a l'entrevistat. El text lliure pot permetre detectar les preguntes rellevants per a un individu en funció dels temes que li interessin.

El model desenvolupat en aquest estudi ha de ser extrapolable a altres dominis, necessitats o objectius.

4 Abast del projecte.

El projecte estarà centrat en les xarxes socials, ja que són una font important de "text lliure"; i en particular en les plataformes de "dating online", que permetran

aconseguir els objectius. En particular, OKCupid.

Aquest estudi es concentra en el que les persones diuen, no si una "estratègia discursiva", entesa com a combinació de "temes", té més o menys èxit.

5 Objectius.

Com a **objectiu principal**, determinar quins són els "temes" rellevants que les persones utilitzen quan es volen presentar com un "producte" atractiu davant dels altres, i si el nombre de temes és finit.

Com a **objectius secundaris**:

1. determinar quines són les diferents "estratègies discursives", enteses com a combinacions de temes.

2. analitzar com les diferents "estratègies discursives" són utilitzades per als diferents "perfils" de persones. La definició de persona i els diferents perfils seran desenvolupats al punt "dades".

6 Estat de la qüestió.

Mesurar la compatibilitat que té una parella potencial abans d'invertir temps, esforç i energia emocional en una cita presencial és un objectiu de les persones que participen en les cites online. [Fiore, 2004]

En estudis previs de psicologia de l'atracció en cites online diuen que la fotografia és un predictor fort de l'atractivitat global d'una persona, però no l' únic; la component del text lliure juga un paper important a l'hora de determinar l'atractivitat global d'una parella potencial.[Fiore et al., 2008]

Encara que els usuaris de cites online tenen limitats els senyals que poden emetre per a l'auto-presentació (nom d'usuari, fotografia, text lliure, etc), poden invertir temps creant i revisant els seus perfils per a generar la seva millor autopresentació. Els usuaris, al redactar els seus perfils, tenen motius de competitivitat: primer, per autopresentar-se tan atractius com sigui possible; i segon, per auto-presentar-se de la manera més acurada possible per tal que aquelles persones que siguin considerades atractives online segueixin sent considerades atractives en una eventual cita offline. De totes maneres, els usuaris de les plataformes de cites online tendeixen a un cert nivell d'exageració.[Ellison N., 2006]

A mesura que la capacitat d'anàlisi de textos augmenta, es comencen a fer estudis sobre el llenguatge natural utilitzant perfils de plataformes de cites online. Per exemple, sobre dades procedents de "Yahoo Personals" amb anàlisis de tipus factorial i cluster. [Meenakshi Nagarajan, 2009].

A través de Kaggle és possible accedir a un dataset públic de OKCupid amb els perfils de 59.946 persones. Els primers estudis que es fan sobre aquest dataset són de tipus estadístic que permeten descriure la població i fan anàlisis de freqüències de paraules diferenciant entre homes i dones. Utilitzen un model logístic per fer prediccions del sexe en funció de l'alçada. [Albert Y. Kim, 2015]

Cal esmentar que al maig de 2016, un equip d'investigadors danesos va posar a disposició del públic el projecte "OkCupid dataset", que contenia (a maig de 2016) 2.620 variables que descrivien 68.371 usuaris. Van publicar un document sobre les correlacions entre la religiositat, l'interès polític i la seva participació [Kirkegaard and Bjerrekær, 2016]. Aquest projecte va acabar amb una investigació de la Danish Data Protection Authority.

A mesura que augmenta la facilitat d'anàlisi del llenguatge natural (NLP) apareixen nous tipus d'estudis basats en el dataset OKCupid. Entre d'altres:

Utilitzant una combinació de NLP i "machine learning" es demostra que els individus transmeten informació no desitjada (inconscient) a través dels "textos lliures" autobiogràfics [Shishido et al., 2016]. Els autors analitzen el consum de drogues. A través d'una reducció de categories fan tres grups: els que afirmen que en consumeixen, els que neguen el seu consum, i els que no contesten a la pregunta. Els autors classifiquen els que no contesten a la pregunta amb la informació facilitada en el text lliure. Utilitzen mètodes de processament com Tf-IDF, Log-Odds-Ratio, Non-negative factorization, permutation test i entrenen una regressió logística.

A diferència d'aquest últim estudi, el que no s'ha fet i es vol analitzar és quina és la informació que l'usuari revela de si mateix de forma "conscient", amb l'objectiu de presentar-se com a producte atractiu.

7 Les dades.

En aquest punt es descriuran les dades del dataset de Kaggle OKCupid disponibles per aconseguir els objectius d'aquest estudi (`https://www.kaggle.com/andrewmvd/okcupid-profiles`).

Aquest dataset disposa dels perfils de 59.946 usuaris, amb dades de l'any 2012, residents a una distància de 25 milles de San Francisco, que inclouen informació de 31 característiques. Aquestes característiques es poden agrupar en tres grups:

1. Característiques de classificació: age, sex, status, orientation, ethnicity, height, location, sign, etc.
2. Característiques d'estil de vida: diet, drinks, smokes, drugs, body type,

education, income, job, pets, offspring, religion, speaks, etc.

3. Text lliure:
   (a) My self summary.
   (b) What I'm doing with my life.
   (c) I'm really good at.
   (d) The first thing people usually notice about me.
   (e) Favorite books movies show music and food.
   (f) The six things I could never do without.
   (g) I spend a lot of time thinking about.
   (h) On a typical Friday night I am.
   (i) The most private thing I am willing to admit.
   (j) You should message me if.

Totes les "característiques" i els seus "valors" estan en la llegua original del dataset, en aquest cas l'anglès, i s'anomenen així al llarg de tot l'estudi.

Aquest dataset no disposa de cap variable que permeti mesurar el nivell d'èxit d'un perfil (número d' interaccions amb diferents persones, número de cites offline assolides, etc.), ni informació del xat que els participants fan servir.

En aquest treball s'entendrà per a **persona** cadascun dels diferents usuaris, i es determinaran els **perfils** en funció de les característiques o combinació de característiques descrites anteriorment. L'elevada dimensionalitat del dataset permetrà disposar d'un elevat nombre de perfils diferents.

Per a ser conscients de les possibilitats i tenir una idea de magnituds, a continuació s'exposen algunes dades rellevants:

El dataset disposa de 59.946 usuaris i es descompon en un 40% de "females" i un 60% de "males". (figura 1)

Respecte a l'estatus, 93% són "singles" (només un 0.05% són "married"), i no hi ha gaire diferència entre "females" i "males". (figura 2)

Respecte a l'orientació, el 86% es declaren "straight", un 9% "gay" i un 5% "bisexual". Proporcionalment hi ha més "gay" entre els "males" i més "bisexual" entre les "females". (figura 3)

Respecte a l'edat. Fins als 50 anys, la proporció de 'males" supera les "females". Després, la proporció s'inverteix. (figura 4)

Figura 1: Dataset per sexes.

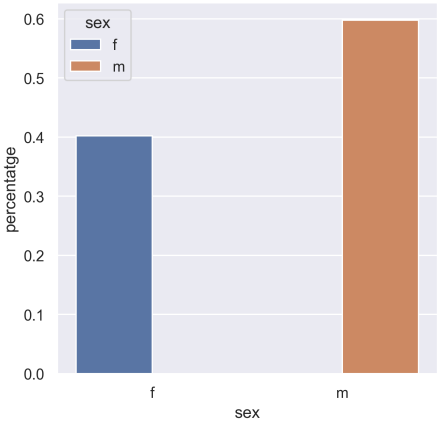


Figura 2: Dataset per status.

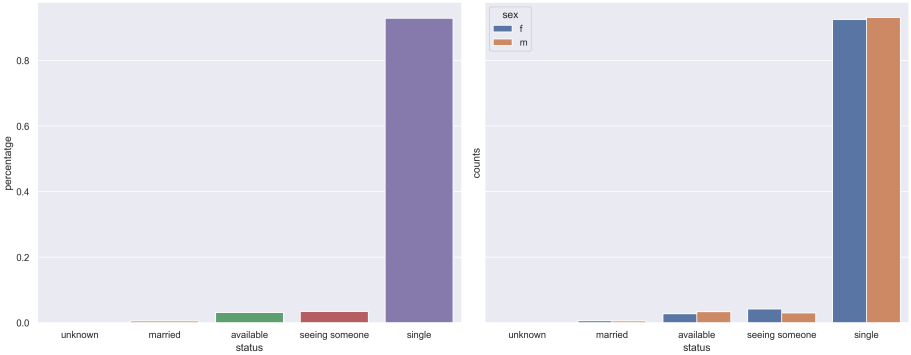


Figura 3: Dataset per orientació.

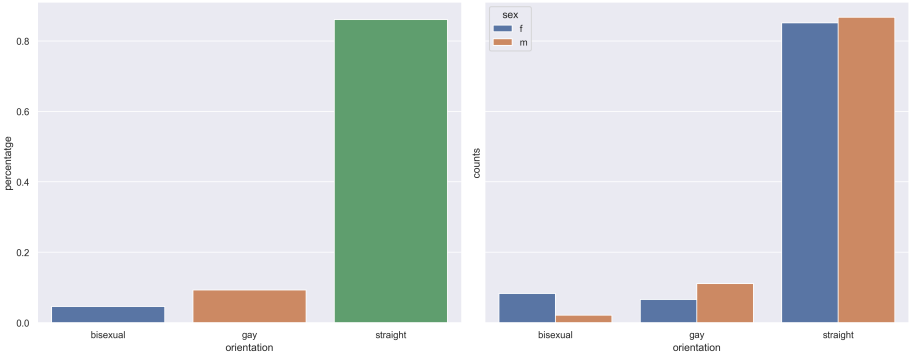
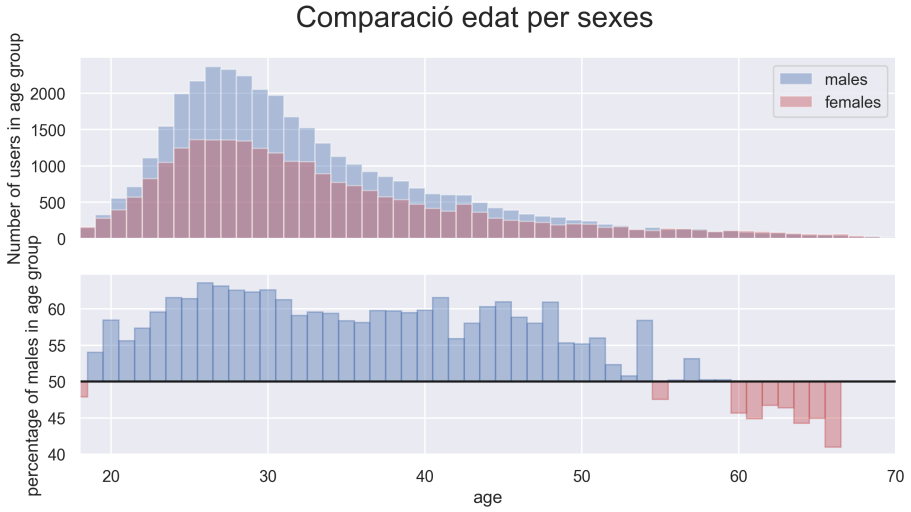


Figura 4: Dataset per edat i sexe.



8 Aspectes de disseny i desenvolupament.

Es parteix del dataset de OKCupid amb els quasi 60.000 perfils.

L'objectiu principal és descobrir quants i quins són els "temes" rellevants que els usuaris fan servir per a presentar-se com a "producte" atractiu. Aquest objectiu s'assolirà amb les característiques de 'text lliure". Com a tècnica per extraure les "paraules" rellevants s'utilitzarà Tf-idf i per a "temes" i "paraules clau" LDA-Mallet.

Com es podrà comprovar, no tots els individus utilitzen els mateixos temes per a presentar-se com a producte atractiu. El que realment fan és utilitzar combinacions de "temes" rellevants per a ells amb diferents intensitats, uns es concentren en uns temes, i uns altres en altres. Caldrà trobar una manera per determinar quan un "tema" és rellevant o no en el "discurs" d'una persona. La idea és modelitzar els "patrons" de combinacions de "temes" rellevants, als que denominarem "estratègies discursives". Es farà servir anàlisis clúster per a determinar aquests patrons. Això permetrà aconseguir el primer objectiu secundari.

Per a determinar els principals "temes" i les diferents "estratègies discursives" es farà servir tot el dataset en el seu conjunt.

Per analitzar si existeixen diferents "estratègies discursives" entre els diferents "perfils" de persones es faran servir les característiques de classificació i estil de vida ("age", "sex", "status", "orientation", etc.). Això permetrà aconseguir el

segon objectiu secundari.

9 Experiments, validació i resultats.

9.1 Selecció de textos lliures rellevants per a la investigació.

Tal com s'ha vist existeixen 10 variables de "text lliure" i tots es podrien considerar com a rellevants. Però per aconseguir els objectius aquest estudi es centrarà en "My self summary".

La principal raó és que "My self summary" és més general, més "lliure", per això el que es podria considerar com el més representatiu per aquest estudi; les altres variables fan referència a aspectes més concrets.

Com a raó secundària, en les altres variables disminueix la riquesa del text aportat (menys paraules per variable) i augmenta el volum d'individus sense resposta (nuls) a mesura que l'usuari avança escrivint els diferents textos (figura 5).

Per una altra part també s'ha intentat treballar amb tots els textos junts, però dificulta l'exposició dels resultats i es corre el risc que els "temes" detectats es corresponguin amb cada una de les altres variables.

9.2 Dels "discursos" a les "paraules": Preprocessament

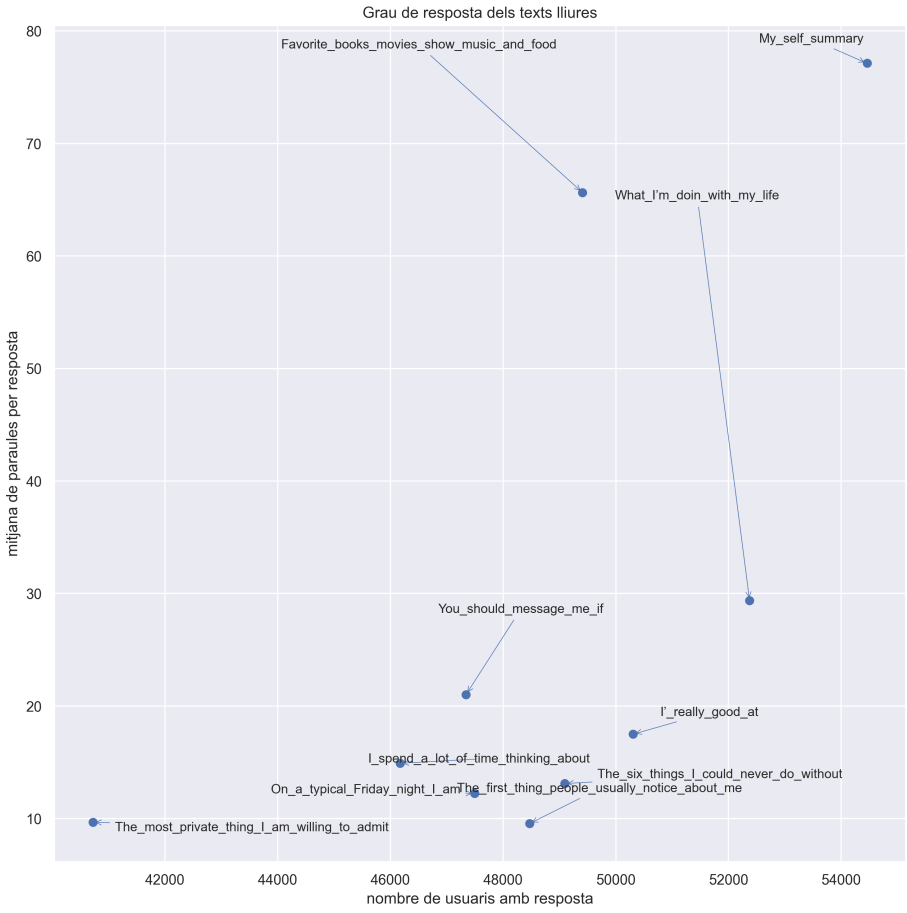
Per aconseguir els "termes" hauria anat bé utilitzar l'extracció de sintagmes nominals com "named entity" de Spacy, però és una estratègia "depenent de la llengua". S'ha preferit utilitzar tècniques d'extracció de termes "independents de la llengua" (amb la idea de poder utilitzar la metodologia en altres idiomes) com els tokenitzadors, bigrams i trigrams de la llibreria de gensim (versió 3.8.3). En el cas dels lemes i els stopwords no és possible utilitzar tècniques independents de la llengua i s'ha fet servir Spacy (versió 2.3.2) i NTLK (versió 3.5), respectivament.

La preparació de les dades per determinar les paraules ha consistit en 4 passos:

1. **tokenitzar**, mitjançant la funció "simple_preprocess" de gensim.
2. **bigrams i trigrams** amb les funcions Phrases i Phraser de gensim. A continuació exposo alguns del resultats obtinguts:

   Electrical engineering, scuba dive, steve job, bar hopping, half moon bay, redwood trees, early twenties, comic books, single malt scotch, laundry list, art gallerie, turn off, daily routine, graduate student, tel aviv, thrift stores, olive oil, cirque soleil, lemmon water, season ticket, medium sized, liberal

Figura 5: Grau de resposta i riquesa del text lliure.



art, foreign language, dating site, bottom line, hablo espanol, fantasy novels, princess bride, open minded, brad pitt, older brothers, give me a sec, clean sheets, self review, shoulder cry, love arielle, mid august, wall street, smarty pant, wear flip flop, twin sister, firm believer, visual art, middle eastern, positive outlook, cultural events, peace corps volunteer, outdoor activities, rock star, mario kart, web develop, grocery shopping, computer programmer, advertising agency, fair warning, fantasy novel, russian hill, goal orientate, self promotion, kindred spirit, secret agent.

3. **lematitzar** tots el tokens que eren NOUN, ADJ, VERB, ADV i PROPN amb la llibreria spacy
4. **stopwords** amb la funció stopwords de nltk.

9.3 Detecció de "paraules rellevants" (Tf-idf).

Com a primera aproximació es vol determinar quines són les paraules rellevants del text lliure "My self summary" calculant el Tf-idf. Per fer-ho, s'ha comparat amb els altres 9 textos lliures junts. El resultat són dos documents, un amb "My self summary" i un altre amb"els altres".

Amb aquests dos documents, s'ha creat un text comú amb el qual s'ha generat el Dictionary (llista de tots els termes, en total 140.462) i el Corpus (que consisteix a convertir els documents a "bag of words"). Després s'ha calculat el Tf-idf. Les 10 paraules més rellevants, tenen tf-idf entre 0.15 i 0.39, i són:

Love, life, people, good, time, thing, new, friend, enjoy i work.

El resultats no són gaire espectaculars, la principal raó és que ens falta context per a interpretar aquestes paraules. Aquest és el motiu pel qual es farà servir LDA juntament amb Word2Vec per a determinar els temes.

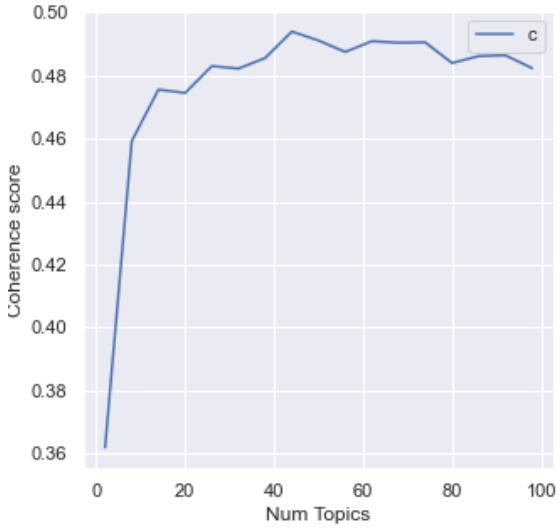
9.4 Detecció dels "temes" i les seves "paraules clau". (LDA)

El "modelat de temes" és un objectiu important en el processament del llenguatge natural. N'hi ha de dos tipus: "predictius" de la categoria d'un document; i els que determinen els "propis temes" tractats en els diferents documents.

Aquest estudi es basarà en aquest segon tipus, els que determinen els propis temes. Els temes queden descrits com a una llista de paraules clau. La metodologia consisteix a presentar aquestes llistes de paraules clau a un humà per a "etiquetar" el significat del tema, que en última instància és subjectiu.

Un dels principals problemes radica en avaluar la qualitat dels temes des de

Figura 6: Coherence score.



la perspectiva de la interpretabilitat humana. Una manera d'avaluar-la és la "coherència semàntica" o "coherence score" que mesura la proximitat semàntica entre les "paraules clau" d'un tema. Quan més pròximes siguin les paraules clau de cada tema, més coherent seran els temes [Morstatter and Liu, 2017].

LDA (Latent Dirichlet Allocation) és una tècnica per a la detecció de temes i serà la que es farà servir en aquest estudi. S'utilitzarà la implementació Mallet de gensim, que permetrà obtenir millors resultats que el simple LDA.

1. **Passos previs**: Creació del Dictionary i el Corpus amb Gensim. S'ha netejat aquells tokens que no hagin estat utilitzats per un mínim de 10 usuaris i que hagin estat utilitzats per més del 50% dels usuaris.

2. S'ha determinat el **nombre de temes** amb el "coeficient de coherència" abans referenciat. A més valor del coeficient de coherència, millor. Després de diferents anàlisis s'ha optat per 7 temes en comptes dels 18 o inclús 42 que recomanaria el gràfic de coherència (figura 6), ja que no representa una pèrdua important d'aquest coeficient i simplifica l'exposició dels resultats.

3. Generació dels **temes i paraules** clau de cada tema. Això es fa buscant les paraules més probables de cada tema. És un resultat del model LDA Mallet

9.5 "Etiquetatge" dels "temes". (Word2Vec)

Per poder etiquetar els temes, s'ha fet en tres fases:

1. **Visualitzar les "paraules clau" de cada "tema"** S'ha fet servir dues tècniques diferents, que es complementen entre elles:

   (a) **Wordcloud** dels diferents temes, que dóna una visualització d'importància relativa de les paraules clau dintre d'un tema. (figura 7).

   (b) **pyLDAavis** permet també visualitzar la rellevància dels termes dintre d'un tema i apreciar les diferències entre els temes (figura 8)

2. **Visualitzar les "paraules clau" de cada "tema" tenint en compte el seu context**. Els dos sistemes de visualització anterior han permès determinar la importància relativa de les paraules basades en la seva freqüència sense tenir en compte la proximitat semàntica. Això s'aconsegueix fent servir **Word2vec** de gensim amb **TSNE** de sklearn (figura 9).

   (a) Word2Vec permet representar les paraules com a vectors en un espai multidimensional, de forma que les paraules semblants semànticament estiguin representades per punts pròxims, tenint en compte el context de les paraules adjacents. S'ha fet servir Word2Vec sobre tots els "discursos" de "My self summary", treballat amb vectors de 150 dimensions, amb un context de 10 paraules tant prèvies com posteriors.

   (b) 'Per a visualitzar les "paraules clau" de cada "tema" s'ha reduït les 150 dimensions de cada paraula, obtingudes amb el Word2Vec, a dues amb TSNE i les hem representat al pla. TSNE és una tècnica no lineal, no supervisada, utilitzada principalment per a la visualització de dades d'alta dimensionalitat. Es diferencia de PCA (análisis de components principals) en el fet que aquesta és una reducció de la dimensionalitat de manera lineal que busca maximitzar la variança.

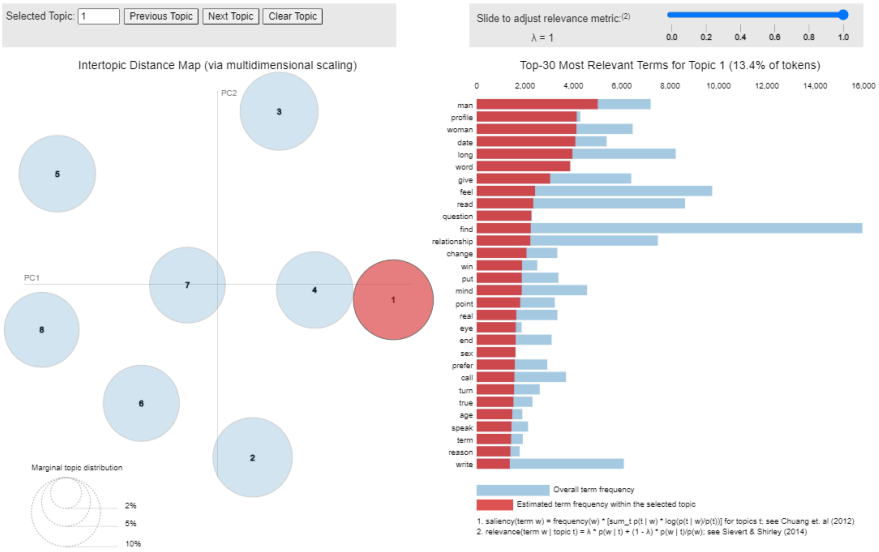
   Tal com s'ha explicat al punt anterior, word2vec més TSNE ajuda a posar en context les paraules que pertanyen a un tema. Per exemple, permet interpretar les paraules: "love", "day", "beach" com: "I love spending a day at the beach" al tòpic1, que juntament amb altres combinacions de paraules permet etiquetar-ho com "Activities I like to do", com es veurà al punt següent (figura 9).

3. **Etiquetar els temes**. A continuació s'ha procedit a etiquetar els temes amb la metodologia explicada anteriorment. La finalitat de l'etiqueta és facilitar la interpretació dels resultats. Els 7 temes detectats són:

   (a) Etiqueta tòpic0: "**Interests and things I do in my free time**". Paraules clau: *'music', 'watch', 'book', 'read', 'make', 'show', 'art', 'drink', 'day', 'stuff', 'dog', 'eat', 'car', 'walk', 'listen', 'coffee', 'house', 'animal', 'game', 'nerd'*(figura 10).

Figura 7: WordCloud.

Figura 8: pyLDAvis.



(b) Etiqueta tòpic1: "**Activities I like to do**". Paraules clau: *'love', 'enjoy', 'travel', 'music', 'food', 'explore', 'live', 'place', 'adventure', 'run', 'great', 'cook', 'city', 'home', 'hike', 'active', 'day', 'beach', 'movie', 'hiking'* (figura 9).

(c) Etiqueta tòpic2: "**Feelings: How I feel about communication and talking with people**". Paraules clau: *'write', 'make', 'profile', 'talk', 'read', 'word', 'give', 'people', 'date', 'long', 'call', 'feel', 'put', 'win', 'bad', 'question', 'message', 'hate', 'end', 'turn'.*

(d) Etiqueta tòpic3: "**How I like to spend my time and what is important to me**". Paraules clau: *'time', 'thing', 'life', 'people', 'good', 'make', 'lot', 'find', 'friend', 'work', 'meet', 'learn', 'hard', 'play', 'family', 'great', 'happy', 'world', 'spend', 'part'.*

(e) Etiqueta tòpic4: "**What I look for in romatic relationships**". Paraules clau: *'life', 'relationship', 'man', 'woman', 'open', 'share', 'world', 'creative', 'heart', 'passionate', 'partner', 'interest', 'feel', 'mind', 'nature', 'important', 'experience', 'interested', 'strong', 'passion'*(figura 11).

(f) Etiqueta tòpic5: "**My backstory and future aspirations**". Paraules clau: *'year', 'live', 'work', 'move', 'grow', 'back', 'bay area', 'play', 'san francisco', 'city', 'bear', 'school', 'college', 'recently', 'raise', 'job', 'ago', 'start', 'spend'* (figura 12).

(g) Etiqueta tòpic6: "**Personality traits**". Paraules clau: *'love', 'good', 'fun', 'friend', 'person', 'guy', 'laugh', 'people', 'pretty', 'enjoy',*

Figura 9: Word2Vec del tema “Activities I like to do“

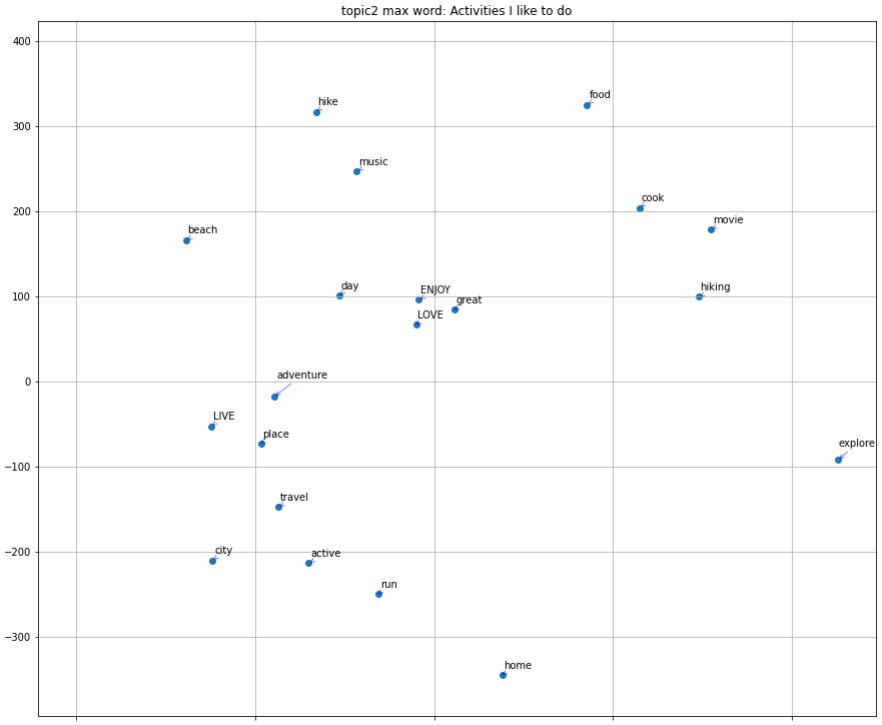
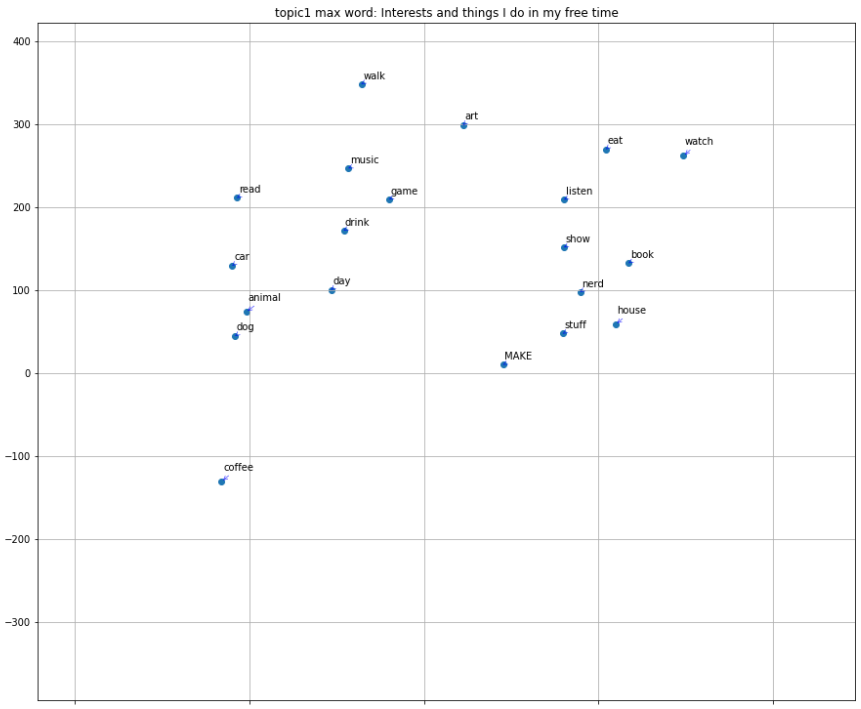


Figura 10: Word2Vec del tema "Interests and things I do in my free time."



'kind', 'meet', 'easy', 'girl', 'nice', 'funny', 'hang', 'sense humor', 'honest', 'thing' (figura 13).

És important esmentar que encara que s'han detectat molts bigrames i trigrames, només "san francisco", "bay area" han aparegut com a paraules clau, en aquest cas per l tòpic5, **"My backstory and future aspirations"** (figura 12); i "sense of humor" al tòpic6 **"Personality traits"** (figura 13).

També s'observa que la paraula "relationship" està pròxima a "heart", "share", "creative", "experience", "open". Per una altra banda, les paraules "interest/interested" estan pròximes a "passion/passionate", "feel", "woman", "man", "partner" al tema **"What I look for in romantic relationships"**.

En **"Personality traits"** les paraules "people/person" estan pròximes a "honest", "easy", "fun/funny/laugh", "kind".

Figura 11: Word2Vec del tema “What I look for in romatic relationships.“

Figura 12: Word2Vec del tema “My backstory and future aspirations“.

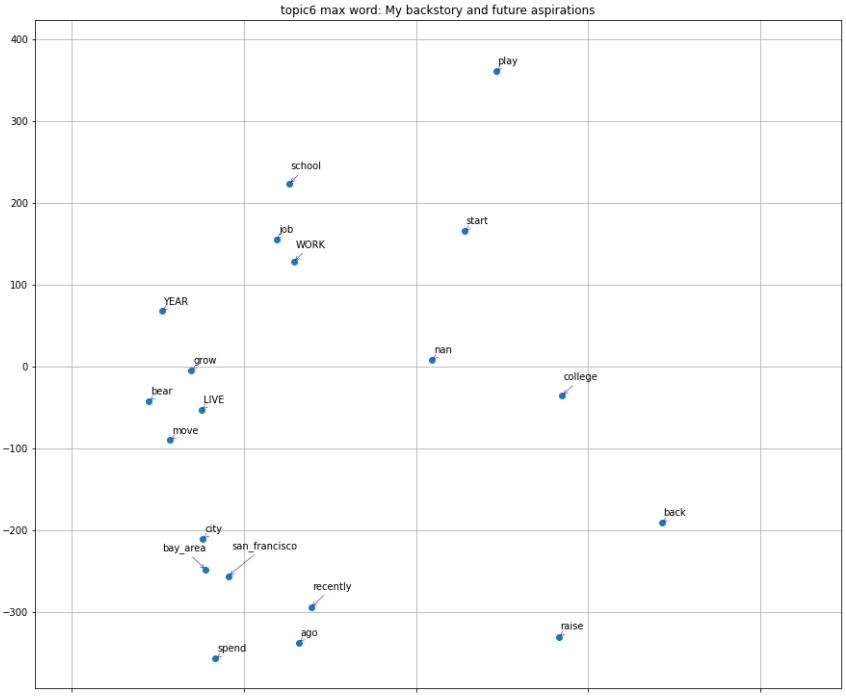


Figura 13: Word2Vec del tema “Personality traits“.

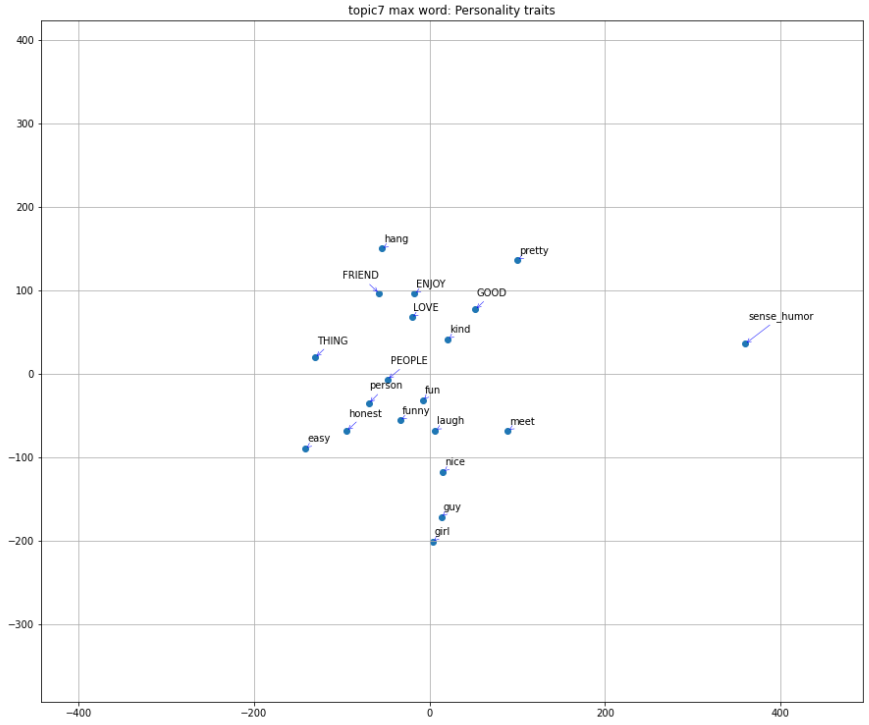
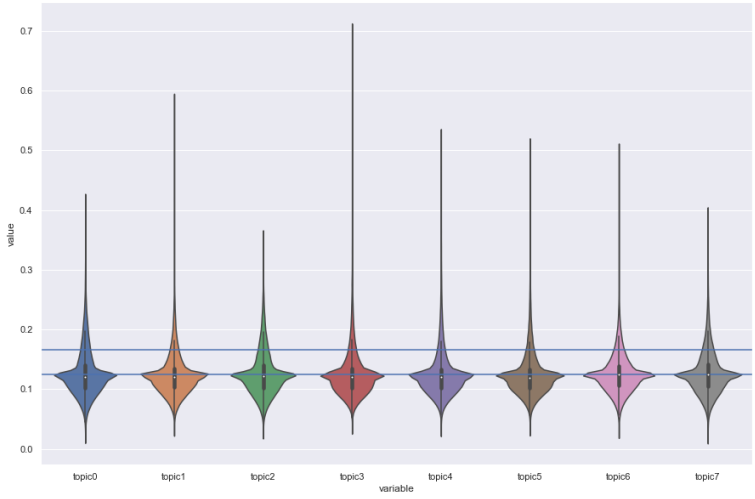


Figura 14: Importancia del temes.



9.6 El "perfil discursiu" d'un individu.

Un usuari en el seu "discurs" fa referència a determinats "temes": a uns els dóna molta importància, a d'altres menys; això fa que els diferents usuaris tinguin "perfils discursius" diferents. En aquest punt es vol assignar a cada usuari el pes que tenen els diferents "temes" en el seu "discurs". Si un usuari fes servir tots el temes en la mateixa proporció, donat que es parla de 7 temes, a cada un d'ell faria servir 14.3%

Això s'aconsegueix aplicant el model LDAMallet al corpus.

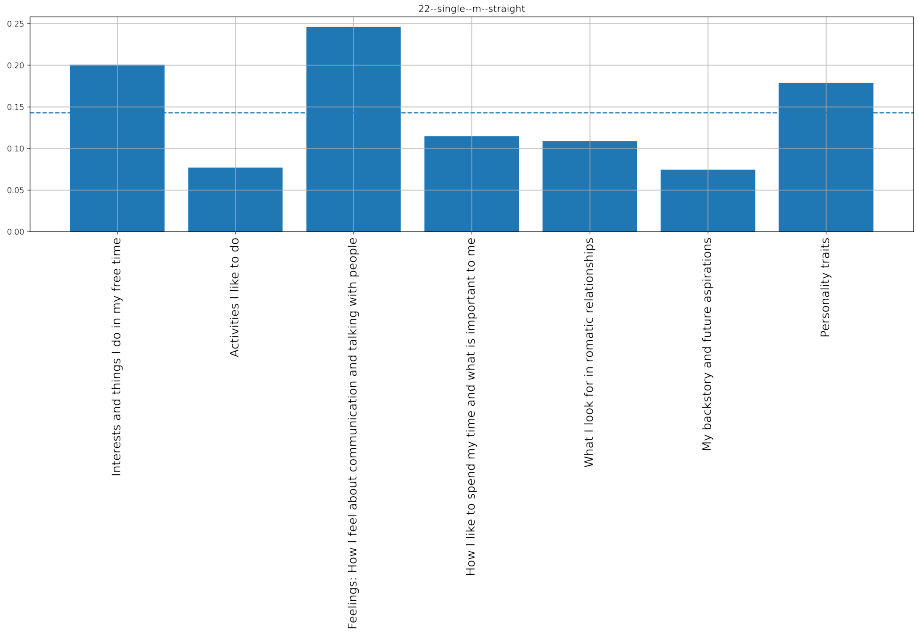
Si s'analitzen les dades a **nivell global**, de tots els usuaris alhora, s'observa allò previst: com a mediana, tots el temes pesen aprox. 14.3%. (figura 14)

Si s'analitzen les dades a **nivell individual** es pot visualitzar el "perfil discursiu" d'un individu.

Si per exemple s'analitza el cas de l'usuari0 ("age": 22, "status": "single", "sex": "male" i "orientation": "straight") (figura 15) s'observa que dedica:

1. Un pes del 24% al tema "**Feelings, communication and talkings**", on apareixen les frases "*to have really serious, really deep conversations about really silly stuff*" o "*I love to talk about ideas and concepts*".
2. El tema "**Interest and things to do in my free time**" representa un

Figura 15: Perfil discursiu de l'usuari 0.



20% del seu discurs amb frases com "*my favorite video game*".

3. El tema "**Personality traits**" amb un pes del 18% amb frases com "*you don't have to be funny, but you have to be able to make me laugh*".

4. També es pot observar que hi ha temes no rellevants en el discurs d'aquest usuari, per exemple "**Activities I like to do**", ja que el seu pes relatiu és molt baix.

És a dir el "perfil discursiu" d'un individu es pot definir com una "combinació lineal" de temes, en el qual els pesos representen la rellevància del tema.

[0.20, 0.08, 0.25, 0.11, 0.11, 0.07, 0.18]

9.7 Determinació dels "temes" rellevants d'un "perfil discursiu".

En el "discurs" de cada individu, s'ha de trobar un mecanisme per determinar quins són realment els temes rellevants del discurs d'un individu. S'han fet tres aproximacions:

1. **Normalització horitzontal** dels pesos dels temes: Es posa un 1 al **"tema"** més rellevant d'un usuari i un 0 al tema menys rellevant, en el seu "perfil discursiu". La resta de temes s'escalen.

2. **Normalització vertical** dels pesos dels temes: Es posa un 1 a l'**"usuari"** que per a un tema en concret és el més rellevant i un 0 per l'usuari que el tema és menys rellevant. La resta d'usuaris s'escalen.

3. Determinar com a rellevants aquells temes que excedeixin en una **desviació estàndard** la mitjana del pes del tema.

Analitzats els resultats de cada aproximació, s'ha optat pel mètode de la mitjana més una desviació estàndard.

Això ha permès considerar el discurs d'un **individu** com un vector, amb un 1 per als temes rellevants per a ell i un 0 per als temes que no ha utilitzat o no han estat rellevants. En el cas el "perfil discursiu" de l'usuari0 passaria a ser:

[1, 0, 1, 0, 0, 0, 1]

Aquest tipus de vector és el que s'utilitzarà per a determinar les "estratègies discursives".

Per una altra banda, si es comptabilitza el número d'1s de cada tema, a nivell **global**, és a dir per a tots els individus, es pot determinar la importància dels temes per als individus quan es presenten com a "producte" atractiu.

Així, el tema "**Personality traits**" el tracten com a important 8.841 usuaris, i el "**Activities I like to do**", 8.380. És a dir, els temes més rellevants tenen les dues components: la "psicològica" (quins són els trets rellevants) i la d' "activitats" (quines són les activitats que m'agraden)(figura 16).

9.8 Determinació de les "estratègies discursives".

Un cop determinat en el punt anterior quins són els "temes" rellevants en el discurs de cada individu, es busquen quines són les diferents "estratègies discursives". Tal com han estat definides, considerarem "estratègies discursives" **patrons** de combinacions de temes importants.

S'aplica la tècnica K-means sobre els "perfils discursius" de tots els usuaris per trobar aquests patrons. Es detecta per la "elbow rule" que es poden classificar en 8 "estratègies discursives" diferents (clusters) (figura 17).

Els principals resultats són (figura 18):

1. Una "estratègia discursiva" es basa en una combinació de temes.

2. Totes les "estratègies discursives" tenen un tema principal combinat amb altres temes.

3. Totes les "estratègies discursives" tenen temes que no són utilitzats.

Figura 16: Rellevància global dels temes per a tots els individus

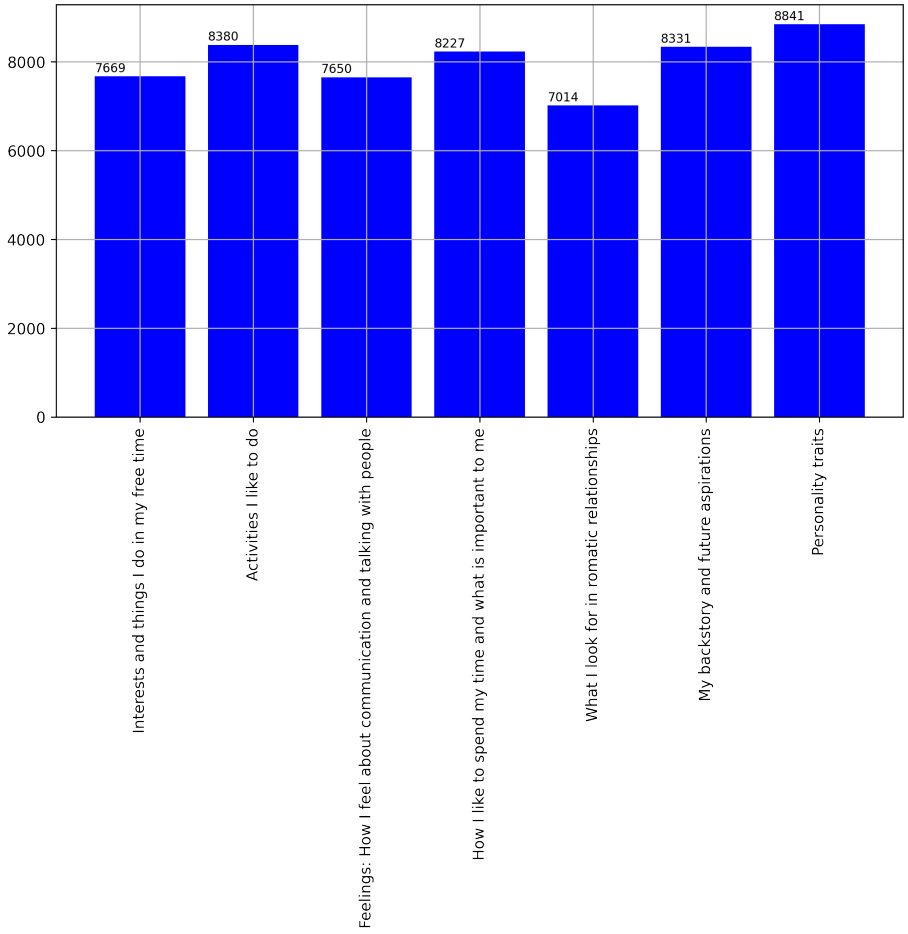


Figura 17: Elbow rule dels clusters d'estratègies discursives

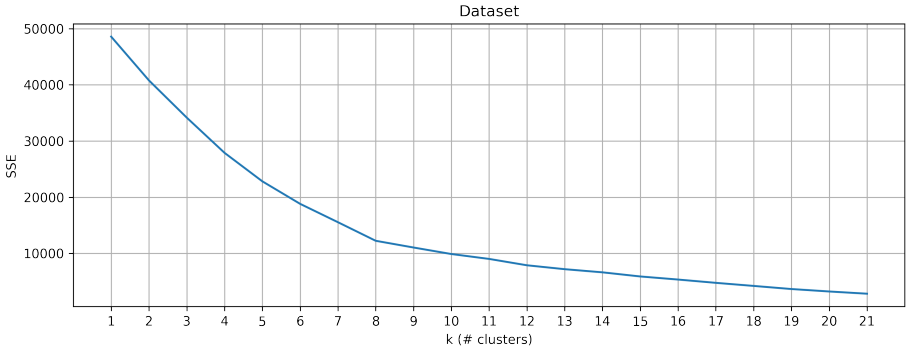
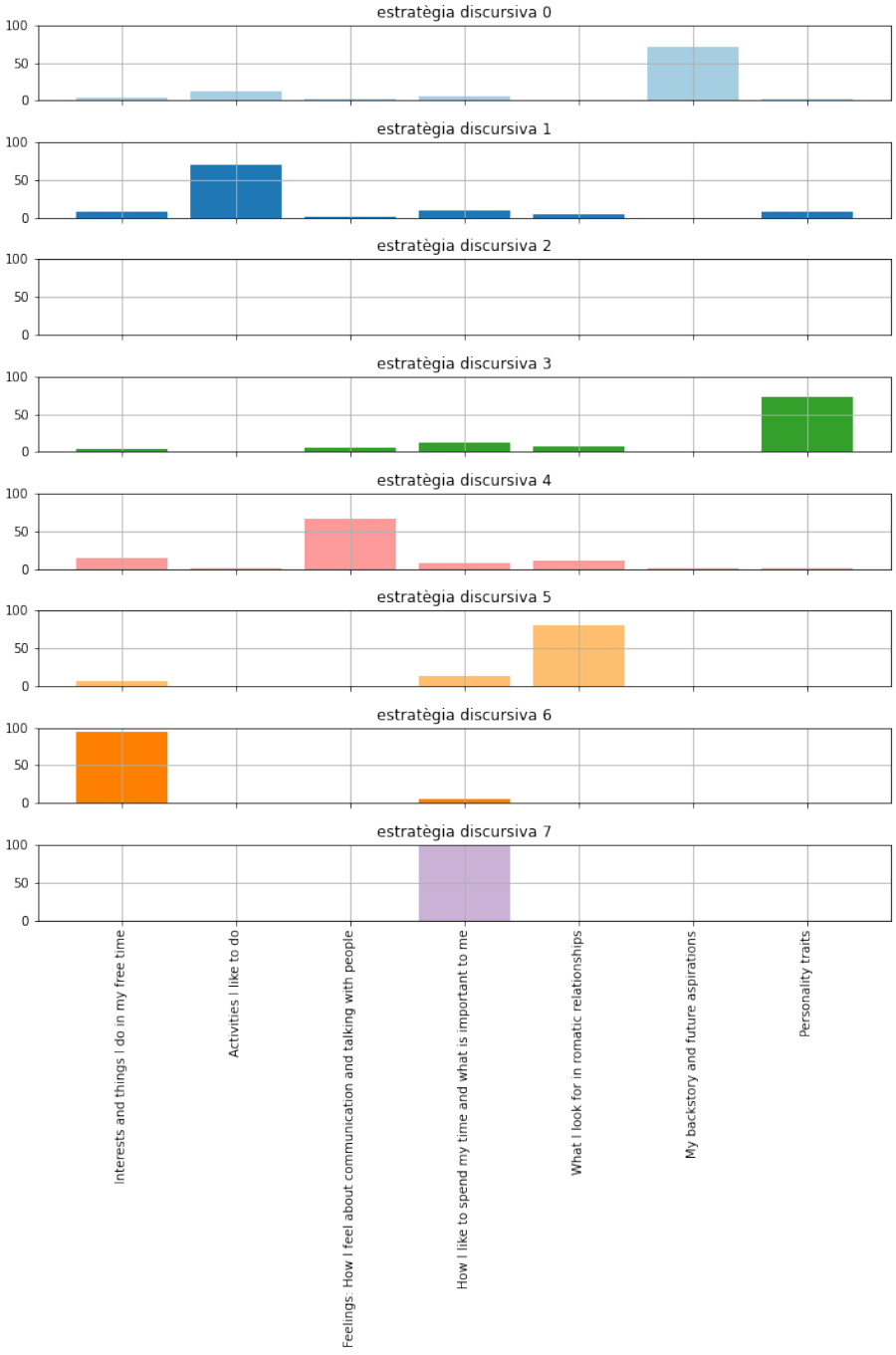


Figura 18: Estrategies discursives



4. Existeix una "estratègia discursiva" que consisteix a no contestar a la pregunta.

5. Només hi ha una "estratègia discursiva pura", és a dir que només utilitza un únic tema.

Definint les diferents "estratègies discursives" i la seva rellevància en funció dels usuaris que la fan servir, tindríem:

1. **Estratègia discursiva 0**: Predomina el tema "My backstory and future aspirations". Aproximadament la utilitzen 11.600 usuaris.

2. **Estratègia discursiva 1**: Predomina el tema "Activities I like to do". Uns 10.000 usuaris aproximadament la utilitzen.

3. **Estratègia discursiva 2**: Usuaris sense resposta.

4. **Estratègia discursiva 3**: Predomina el tema "Personality traits" (uns 10.500 usuaris).

5. **Estratègia discursiva 4**: Predomina el tema "Feelings: How I feel about communication and talking with people" (uns 10.000 usuaris).

6. **Estratègia discursiva 5**: Predomina el tema "What I look for in romatic relationships" (uns 5.600 usuaris)

7. **Estratègia discursiva 6**: Predomina el tema "Interests and things I do in my free time"

8. **Estratègia discursiva 7**: Es concentra en "How I like to spend my time and what is important to me" amb uns 3.600 usuaris

Ara que disposem de les diferents "estratègies discursives" podem determinar quina fa servir cada individu. Un cop tenim les dades a nivell individu les podem agrupar a nivell de "perfil". Això ens permetrà determinar les "estratègies discursives" dels diferents "perfils" d'individu ("sex", 'status", etc.)

9.9 Anàlisi entre les "estratègies discursives" de les "tipologies" d'individu.

A continuació es comparen les "estratègies discursives" entre les diferents tipologies d'individus.

Analitzem en primer lloc l'estratègia discursiva en funció de la característica **"sex"**:

1. les dones estan millor representades en l'estratègia discursiva 1: on predomina (tal com hem vist al punt anterior) "Activities I like to do" i la 5: on predomina "What I look for in romatic relationships";

2. mentre els homes tenen una estratègia de tipus 0: en què predomina "My backstory and future aspirations".

Comprovem amb una Chi2 que les estratègies discurves depenen del sexe (p-value= 0.000).

Continuem analitzant les "estratègies discursives" per a altres característiques:

Posant atenció a les **característiques de classificació**:

1. **Per "age"**: A la variable edat se li ha fet una discretització en tres nivells (s'ha fet amb un cluster tipus K-means, segons la "elbow rule" de la variable): "18-31 anys"(34.625 usuaris), "32-45 anys" (19.253 usuaris) i "més de 45 anys (6.068 usuaris).
   (a) El més joves es decanten per l'estratègia tipus 0: en la qual predomina "My backstory and future aspirations" entre altres,
   (b) mentre que les altres edats prefereixen la 5: en la qual predomina "What I look for in romatic relationships".
2. **Per "orientation"**:
   (a) els bisexuals estan més representats en l'estratègia tipus 4: en la qual predomina "Feelings: How I feel about communication and talking with people",
   (b) mentre els gay en el tipus 5 en la qual predomina "What I look for in romatic relationships".
3. **Per "status"**: el 93% dels usuaris declaren que són "single". Els "married" que representen el 0.5% dels usuaris i els "seeing someone" 3.4%, es decanten principalment per l'estratègia discursiva 4 en la qual predomina "Feelings: How I feel about communication and talking with people".

Quant a les **característiques d'estil de vida**:

1. **Segons la manera de beure("drinks")** el 70% és "socially".
   (a) Els "rarely" 9.9% dels usuaris es decanten pel tipus discursiu 5: predomina "What I look for in romatic relationships";
   (b) en canvi, els "socially" ho fan pel 1: amb predomini de "Activities I like to do".
2. **Segons el nivell de formació ("education")**:
   (a) els "college/university" es decanten principalment per l'estratègia discursiva 2: consisteix a no contestar;
   (b) els "graduated from college/university"prefereixen la 1: predomina "Activities I like to do";

(c) mentre que els "graduated from masters program", per la tipus 5: predomina "What I look for in romatic relationships".

Es pot dir que les anteriors característiques són les més rellevants. Encara que també s'han analitzat els resultats per a les altres característiques no s'esmenten en aquest document a fi de no estendre'l.

10 Principals conclusions.

Amb aquest treball s'han aconseguit els objectius que es perseguien:

1. Ha quedat demostrat que el "topic modeling" permet determinar els principals temes que les persones utilitzen per a presentar-se com a "producte" atractiu davant dels altres. També ha quedat demostrat que el nombre de temes és finit, en aquest cas només han calgut 7 temes.

2. També ha quedat demostrat que existeixen diferents "estratègies discursives", enteses com a combinacions de temes rellevants, utilitzades per a presentar-se com a "producte" atractiu.

3. I per últim, que els diferents "perfils" de persones utilitzen diferents "estratègies discursives".

Totes les "paraules", "temes" i "estratègies discursives" que els individus fan servir ho fan perquè consideren que els reportarà èxit. En analitzar els resultats s'ha pogut comprovar que en general s'eviten "paraules" que haurien de ser rellevants per a determinats "temes", que puguin ser considerades controvertides per assolir l'èxit esperat. Per exemple, al tema "Interests and things I do in my free time" (figura 10) es parla més de música, llibres, art... que de videojocs, futbol, bàsquet o beisbol, etc.

Amb aquest estudi s'ha pogut demostrar l'existència de "temes", "estratègies discursives", i que són utilitzades de diferents maneres segons la tipologia d'individus. Seria interessant si es pogués extrapolar a altres plataformes de cites online i a altres tipologies de xarxes socials.

També s'ha pogut comprovar que LDA amb word2vec i TSNE facilita determinar i etiquetar els temes. Encara que Tf-idf permet detectar les paraules més rellevants, no permet interpretar-les en context. El principal avantatge d'incorporar word2vec és que aprofita el context i la proximitat semàntica.

Els bigrames i trigrames ajuden a trobar significat a les paraules, però el seu baix pes en els discursos fa que en general no surtin com a paraules clau dels temes. S'ha vist que només han aparegut: "san francisco", "bay area" i "sense humor".

Aquest treball també ha demostrat la importància del "text lliure" com a **font**

d'informació imparcial, ja que permet que l'individu expressi el que realment és rellevant per a ell.

Des d'un punt de vista personal m'ha permès contraposar aquesta metodologia basada en el "text lliure" a l'estàndard dels "qüestionaris". En els qüestionaris és l'investigador el que decideix què és rellevant i què no. Aquesta manera de fer genera un fenomen pervers ja que "obliga" l'individu a fer valoracions a vegades "forçades" sobre aspectes que no han sigut rellevants per a ell. A més a més, aquestes valoracions "forçades" s'agreguen, amb el mateix pes, amb les dels individus per als quals sí són rellevants. És a dir, si en la "descripció lliure" d'una estància en un hotel, un client no esmenta la neteja dels lavabos, es pot interpretar que estaven "normals", si l'hagués sorprès positivament o negativament ho expressaria; si es fa a través d'un qüestionari s'està obligant l'individu a valorar un aspecte no rellevant per a ell.

A priori sembla que la metodologia desenvolupada per analitzar el "text lliure" en la manera que les persones es presenten com a "producte" atractiu pot fer-se servir en altres dominis, ja sigui per analitzar productes, idees, etc.

Si el "text lliure" pogués anar acompanyat d'alguna variable d'èxit (com per exemple en el cas de plataformes online, del nombre de cites offline; o en el cas d'un producte la valoració del mateix, etc) permetria desenvolupar tot un sistema d'"anàlisi de sentiments". Un cop determinats els "models discursius", a través de "machine learning" es podria modelitzar la "valoració" i determinar quines són les estratègies clau, tant globalment com pel que fa a perfils d'individus.

11 Línies de treball futures.

Per poder generalitzar els resultats d'aquest estudi sobre la manera que les persones es presenten com a "producte" atractiu s'hauria de comparar amb estudis en altres plataformes de cites online i en altres xarxes socials com Linkedin, Twitter, Facebook, etc.

El tema de basar els "qüestionaris" en temes rellevants extrets del text lliure obre tota una línia d'investigació.

12 Bibliografia.

Referències

[Albert Y. Kim, 2015] Albert Y. Kim, A. E.-L. (2015). Okcupid data for introductory statistics and data science courses. *Journal of Statistics Education, Volume 23, Number 2.*

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

[Ellison N., 2006] Ellison N., Heino R., G. J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication 11(2):415–441.*

[Fiore et al., 2008] Fiore, A., Taylor, L., Mendelsohn, G., and Hearst, M. (2008). Assessing attractiveness in online dating profiles. *Conference on Human Factors in Computing Systems - Proceedings*, pages 797–806.

[Fiore, 2004] Fiore, A. R. T. (2004). *Romantic regressions: An analysis of behavior in online dating systems.* PhD thesis, Massachusetts Institute of Technology.

[Jones, 1972] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

[Kirkegaard and Bjerrekær, 2016] Kirkegaard, E. O. W. and Bjerrekær, J. (2016). The okcupid dataset: A very large public dataset of dating site users. *Open Differential Psychology.*

[Meenakshi Nagarajan, 2009] Meenakshi Nagarajan, M. A. H. (2009). An examination of language use in online dating profiles. *Proceedings of the Third International ICWSM Conference.*

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

[Morstatter and Liu, 2017] Morstatter, F. and Liu, H. (2017). In search of coherence and consensus: Measuring the interpretability of statistical topics. *J. Mach. Learn. Res.*, 18:169:1–169:32.

[Shishido et al., 2016] Shishido, J., Narasimhan, J., and Haller, M. (2016). Tell me something i don't know: Analyzing okcupid profiles. *Proceedings of the 15th python in science conference.*

Índex de figures

13 Glosary:

- **Tf-idf** (term frequency - inverse document frequency) intenta mostrar com d'important és una paraula que apareix en un document d'una col·lecció de documents o corpus. La importància augmenta proporcionalment al nombre de vegades que la paraula apareix al document (Tf) i es compensa per al nombre de documents en el corpus que contenen la paraula (idf). El concepte idf va ser desenvolupat per Karen Spärck Jones a l'any 1972 [Jones, 1972]

- **LDA** (latent dirichlet allocation) és un model generatiu que permet que conjunts d'observacions puguin ser explicats per grups no observats. Si les observacions són paraules en documents, pressuposa que cada document és una barreja d'un petit nombre de categories (denominades temes) i l'aparició de cada paraula en un document es deu a una de les categories a les quals el document pertany. [Blei et al., 2003]

- **Word2vec** és un algorisme que utilitza una red neuronal per aprendre les associacions de paraules en un gran corpus de text. Una vegada entrenat el model, representa cada paraula com un vector i permet detectar sinònims. Els valors d'aquests vectors són escollits de tal manera que el cosinus entre ells indica el nivell de similaritat semàntica entre les paraules. [Mikolov et al., 2013]