Processament i anàlisi d'informació textual

Curs 2024/2025

Marc Bracons

1. Tria i preparació del dataset (10%)

A partir de la selecció del dataset i de les tasques realitzades en l'apartat 1 de la pràctica, titulada “Preparació del dataset”, contesta a les següents preguntes:

1. Descriu el dataset triat. Indica l'origen d'aquest, la quantitat de dades que conté, la tipologia de les mateixes i justifica la seva elecció.
   S’ha utilitzat el joc de dades “Large Movie Review Dataset (IMDb Movie Reviews), extret de la pàgina oficial de la Universitat de Stanford (https://ai.stanford.edu/~amaas/data/sentiment/). Originalment conté 50.000 ressenyes de pel·lícules en anglès, totes etiquetades amb sentiment positiu o negatiu. Les mostres estan equilibrades, amb un 50% de representació per cada opció.

   S’ha reduït la mida a 7.500 mostres (mentint la proporció de representació de classes) per tal de tenir un temps d’entrenament més curt i poder fer proves de manera més àgil.

   La motivació principal per triar aquest conjunt de dades està en el fet que és un estàndard molt utilitzat en anàlisi de sentiments i processament de llenguatge natural, amb una bona representativitat d’opinions reals i amb la possibilitat de treballar clarament en una tasca de classificació binària

2. Explica els problemes que has trobat en les dades i les activitats de neteja i preprocessament que has realitzat durant l'etapa de “preparació del dataset”.

   Durant la preparació s’han identificat i tractat els següents problemes:
   - **Presència d’etiquetes HTML i URLs:** netejat buscant patrons de les etiquetes HTML (r'<[^>]+>) i de les URLs (http\S+|www\.\S+).
   - **Mencions i hashtags:** Mitjançant expressions regulats per buscar “@” i “#”
   - **Contraccions:** amb la llibreria *contractions*.
   - **Signes de puntuació innecessaris:** mitjançant el conjunt de caràcters a eliminar i expressions regulars.
   - **Duplicats:** amb la funció *drop_duplicates*.
   - **Normalització de majúscules i minúscules:** amb la funció *lower()*.
   - S’ha mirat si hi ha ressenyes extremadament curtes, però no ha sigut el cas.

2. Obtenció de dades (30%)

A partir de les tasques realitzades en l'apartat 2 de la pràctica, titulada “Obtenció de dades”, contesta a les següents preguntes:

1. Comenta i compara els n-grames i les col·locacions obtinguts mitjançant els diferents mètodes en aquest exercici.
   Primer s’han aplicat les mètriques **PMI** i **Likelihood Ratio** per detectar els n-grames més rellevants al corpus. Després s’ha implementat un model de detecció de frases amb **Gensim** per identificar col·locacions.

   - **PMI:** Al tendir a afavorir seqüències de paraules que apareixen juntes i amb poca freqüència ha resultat en una llista de n-grames rars, associats a contextos molt concrets, com noms propis (donna snartlebutt) o termes poc habituals en l’ús general (coon huntin).
   - **Likelihood Ratio:** Aquí es tendeix a detectar n-grames més freqüents en el discords, però amb associació estadísticament significativa (”low budget”, “special effects”, “real life”, ...). A més, no ha pogut trobar 20 trigrames ja que, després d’aplicar els filtres i els criteris, en quedava un.
   - **Gensim:** Aquest mètode identifica seqüències de paraules que apareixen amb una freqüència alta i les tracta com una unitat (unint-les amb “_”). Amb això s’ha pogut obtenir una llista de frases sense paraules buides ni seqüències por rellevants. (yeah, i_watched, this, mini_series, ...).

2. Analitza els termes obtinguts utilitzant el model Word2Vec, tria els 5 termes (aspectes) més rellevants i comenta els criteris tinguts en compte per a la seva selecció.

   Un cop entrenat el model *Word2Vec*, s’ha pogut obtenir relacions semàntiques entre paraules del corpus. Per exemple, les següents:
   - **story/plot:** son dos termes centrals a gairebé qualsevol ressenya, ja que es fan servir per descriure el fil narratiu i la coherència de l’argument.
   - **characters:** El desenvolupament i profunditat dels personatges és un altre aspecte essencial.
   - **Interesting:** indica el grau de curiositat que desperta la pel·lícula.
   - **acting:** S’utilitza per valorar la qualitat interpretativa els actors.
   - **twist:** Els girs de guió son un tema comú en les ressenyes.

3. Detecció de temes (30%)

A partir de les tasques realitzades en l'apartat 3 de la pràctica, titulada “Detecció de temes”, contesta a les següents preguntes:

1. Analitza els resultats obtinguts en la primera part de l'exercici, l'exploració dels temes amb WordNet.

   S’ha vist que existeixen temes que poden ser similars al model *word2vec*, poden ser tindre una baixa similitud semàntica segons *Wu & Palmer*. Per exemple, el tema “**story**”, en el model *word2vec* té com a tema proper “**plot**” amb una similitud de 0.766 o “**characters**” amb un valor de 0.7105. En canvi, quan avaluem la similitud amb *WordNet*, el valor per “**plot**” es redueix a 0.2667 i per “**character**” no s’ha ni trobat.

   Això ens indica que en *WordNet* es capten associacions contextuals d’ús que poden anar més enllà de les relacions lèxiques clàssiques. També hem vist que no hi apareixen abreviatures com **imdb**, **cgi** o **fi**, que son habituals en les ressenyes de pel·lícules però son termes tècnics.

2. Compara els diferents models LDA utilitzats, tria el més adequat de forma justificada.

   S’han provat tres configuracions diferents de models LDA, amb 5, 10 i 15 temes. Per triar el model òptim s’han fet servir mètriques de coherència i perplexitat.

Model	Coherència	Perplexitat
5	0.449	-8.45
10	0.436	-9.23
15	0.353	-10.29

   El model 5 té el millor equilibri entre interpretabilitat dels temes i qualitat de les mètriques. Tot i que el model amb 10 tòpics té una coherència més alta, la coherència és més baixa, indicant que els temes son menys significatius.

   També s’ha fet servir una eina per veure la distància entre els temes. Per exemple si triem el model amb 10 tòpics podem veure (figura 1) que hi ha temes que estan molt junts (9 i 10), indicant que comparteixen conceptes comuns. Temes més distanciats (7 i 2) estaran més separats semànticament. El model 10 té una distribució equilibrada, amb certa superposició en alguns casos.

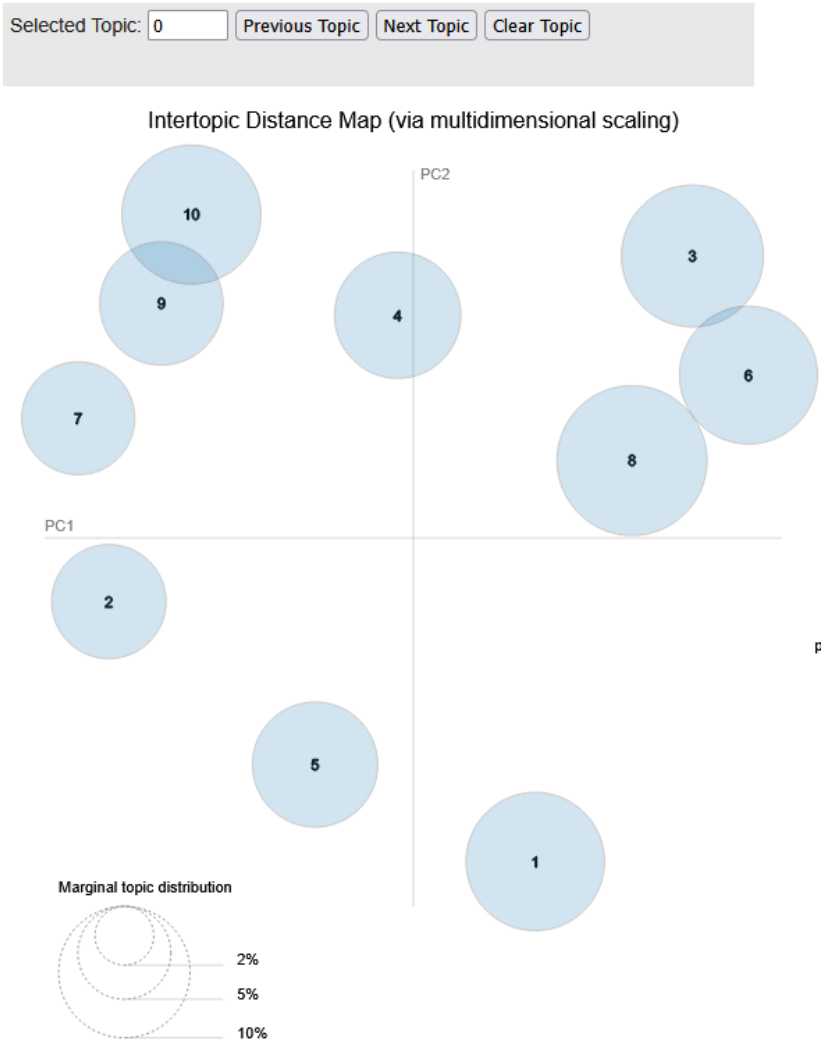

Figura 1: intertopic Distance Map

4. Classificació automàtica d'opinions positives i negatives (20%)

A partir de les tasques realitzades en l'apartat 4 de la pràctica, titulada “Crear un classificador automàtic d'opinions positives i negatives”, comentar els algorismes utilitzats, els resultats obtinguts i la coherència dels resultats amb el contingut dels comentaris.

Pel classificador primer s'ha vectoritzat el text amb **TF-IDF**, transformant les opinions en vectors numèrics, considerant la importància relativa de cada paraula en el conjunt. A més, s'ha fer servir el paràmetre “*word*” per treballar a nivell de paraula individual.

A continuació s'ha entrenat el classificador amb Logistic Regression, un model que estima la probabilitat de pertànyer a una de les dues categories (positiu o negatiu). Aquesta elecció és adequada per a una classificació binària, com en el nostre cas.

Això ha permès classificar i obtenir una interpretació de les paraules que contribueixen més a cada classe. Per exemple, per a la classe positiva, les paraules que aporten més informació son *great*, *excellent*, best, well, etc; i per a la negativa, *bad, worst, awful, etc*.

Finalment, s'ha fet servir el diccionari **AFINN-111**, per calcular la puntuació mitjana de la polaritat de cada opinió. El model ha pogut detectar tres ressenyes en las que la opinió és molt negativa.

5. Avaluació (10%)

A partir de les mètriques calculades en els classificadors de l'apartat 5 de la pràctica, titulada “Avaluació”, hauràs de comparar i avaluar els dos models proposats en funció de les mètriques d'avaluació vistes a l'assignatura (*precision*, *recall* i f1) i del temps d'execució. Conclou finalment quin dels dos models triaries per a predir noves ressenyes sobre productes i el perquè.

Els resultats dels dos models es poden veure en la següent taula:

Mètrica	Regressió Logística	SVM
Temps d'entrenament	1.28 segons	23.44 segons
Precisió (negativa)	0.87	0.88
Precisió (positiva)	0.86	0.86
Recall (negativa)	0.84	0.85
Recall (positiva)	0.88	0.89
F1-score (negativa)	0.86	0.86
F1-score (positiva)	0.87	0.88
Precisió (global)	0.86	0.87

El model SVM té uns resultats molt lleugerament millor en la majoria de mètriques, però té un cost computacional molt més elevat, de l'ordre d'unes 20 vegades superior. Així doncs, tot i que SVM té un millor rendiment, la diferència és mínima respecta al model de regressió logística, per tant, optaria per aquest últim.