Facultat d'Informàtica de Barcelona

GIA UPC

Processament del llenguatge humà

Pràctica 1

Alumnes :
Casanovas Poirier, Anna
Pumares Benaiges, Irene

Tutors :
Turmo Borrás, Jordi
Medina Herrera, Salvador

April 12, 2024

Contents

Abstract

L'objectiu d'aquesta pràctica ha estat implementar un identificador d'idioma per a 6 llengües: el castellà, l'anglès, el francès, l'alemà, l'italià i el neerlandès. S'utilitzen com a dades 30 mil frases de cada idioma com conjunt d'entrenament i 10 mil de cada idioma com a conjunt de proba.

Per assolir l'objectiu primer es fa un preprocessament de les dades, i després es fa un ús de trigrames de caràcters com a model del llenguatge (també aplicant tècniques de suavitzat).

El model també és evaluat mitjançant mesures com la precisió o la matriu de confusió. Finalment, es fa una anàlisi dels resultats obtinguts i es proporcionen unes conclusions sobre el funcionament del model i possibles millores.

1 Introducció

1.1 Motivació del treball

Durant les primeres sessions de l'assignatura de Processament del Llenguatge s'ha parlat de l'estructura d'un document, que implica temes com la tokenització o la separació d'oracions. Aquest tema ha estat relacionat amb l'identificació d'idioma, en la que shan centrat les classes teòriques i pràctiques.

Aprendre a fer un identificador d'idioma té clarament aplicacions directes amb la realitat, ja que la detecció d'idiomes és utilitzada en molts àmbits actualment com a la traducció automàtica, l'anàlisi de sentiments en xarxes socials, l'identificació de spam en correu electrònic i la personalització de continguts en plataformes digitals.

1.2 Descripció de la tasca i les dades

La tasca que es requereix a la pràctica és fer un identificador d'idioma amb 6 llengües europees: el castellà, l'anglès, el francès, l'alemà, l'italià i el neerlandès. Se'ns ha proporcionat una serie de documents amb moltes frases de cada idioma. En total 40 mil de cada llengüa, però s'utilitzaràn 30 mil per entrenar el model i 10 mil per fer el test.

Aquestes dades provenen de Leipzig Corpora Collection. Contenen frases dels 6 idiomes dividides amb salts de línia.

És necessari fer un preprocessament de les dades, fer l'entrenament del model es farà mitjançant trigrames de caràcters i finalment s'ha de realitzar un anàlisi complet dels resultats.

2 Preprocessament

El preprocessament del text abans d'entrenar i analitzar és una tasca molt important per a la millora del rendiment del model.

Per a fer tot aquest preprocès, s'ha creat una funció anomenada preprocess que, donat un text, retornà el text preprocessat. És important destacar que s'ha realitzat un preprocessament a les dades d'entrenament, però també a les de proba després.

En primer lloc, s'eliminaran tots els digits i els signes de puntuació, ja que no aporten informació rellevant sobre el contingut del text per a identificar l'idioma. D'aquesta manera, el model es centrarà en les lletres, que són representatives de l'idioma. Un altre pas que es seguirà és convertir tot el text a minúscula, ja que ajuda a reduir la complexitat i normalitzar el format del text, el que facilitarà la feina del model. De la mateixa manera, els espais en blanc adiccionals no aporten informació i podrien afectar el rendiment introduint soroll innecesari.

Per últim, en les dades del train, es substituiran els salts de línia per espais dobles, amb l'objectiu de preservar la relació entre el final d'una frase i el principi de la següent. D'aquesta manera, al fer la divisió en trigrames, la separació de frases quedarpa ben indicada. En el cas del test, no hem hagut de fer això ja que s'ha creat una llista i s'ha dividit el text per frases. Això és degut a que la identificació de frase es fa una per una, tot i que entrenem el model amb un text sencer.

3 Obtenció dels trigrames

Per realitzar l'obtenció dels trigrames, s'ha fet servir la funció TrigramCollocationFinder de la llibreria nltk. Aquesta el que fa es retornar una llista de tuples, on tenim el trigrama trobat i el nombre de cops que s'ha trobat al text. Com ens demanava l'enunciat de la pràctica, s'han eliminat els trigrames que tenen menys de 5 aparicions. Això es fa per vàries raons. Primer de tot, els trigrames que són poc freqüents no ens aportaran informació per trobar patrons lingüístics significatius, el que podria proporcionar un overfitting degut al soroll. El model serà més capaç de generalitzar ja que no tindrà en compte característiques poc informatives.

4 Entrenament del model

Per a entrenar el model, es disposa de sis documents de text, un per a cada llengua, amb 30.000 frases cadascun. Després de preprocessar cada document, d'ha decidit implementar un diccionari que emmagatzema els trigrames i les probabilitats corresponents (calculades a la funció probabilitat_suavitzada), un per a cada idioma. Cada probabilitat es calcula mitjançant la fórmula de la Ley de Lidstone, que és una técnica de smoothing que ajusta les freqüències observades per evitar divisiones per zero i per poder treballar amb trigrames que no apareixen a les dades d'entrenament. La fórmula és la següent:

$$P^T(ej) = \frac{C_T(ej) + \lambda}{N_T + B\lambda}$$

· $C_T(ej)$ és el nombre de vegades que apareix el trigrama al corpus
· λ és el parámetre de suavitzat
· N_T és el nombre total de trigrames en el corpus
· B és el nombre total de trigrames diferents

Per a triar el valor de B, s'ha considerat utilitzar un valor diferent per a cada llengua, amb el nombre de lletres elevat a 3 per a cada idioma, però s'ha observat un millor rendiment del model utilitzant la mateixa B per a tots, és a dir, amb la mitjana de B de cada idioma.

Per a l'elecció del valor de λ s'ha utilitzat la validació creuada. S'ha considerat provar valors en l'interval de 0 a 1. S'ha provat amb els següents valors: 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0,.9 i 1. Els valors de 0.05 i 0.1 han proporcionat els millors resultats en termes de precisió.

Finalment s'ha decidit utlitzar B = 29.5^3 (mitjana) i $\lambda = 0.1$ per a aquest model.

$$P^T(ej) = \frac{C_T(ej) + 0.1}{N_T + 29.5^3 * 0.1}$$

5 Anàlisi dels resultats

5.1 Experiment 1

Abans de tot, s'ha provat d'executar el codi amb valor de lambda 1, i amb la B diferent per a cada llengua. Tampoc s'ha tingut en compte l'eliminació dels signes de puntuació en aquest experiment. Al executar, veiem que el percentatge de precisió és ja bastant elevat, amb un valor de 98.0058%. Això vol dir el que model ja és bastant eficaç sense tenir en compte aquests tres factors, però, si triem un valor de lambda millor, la mateixa B per a tots els idiomes i prescindim dels signes de puntuació, el model podria tenir un cert marge de millora.

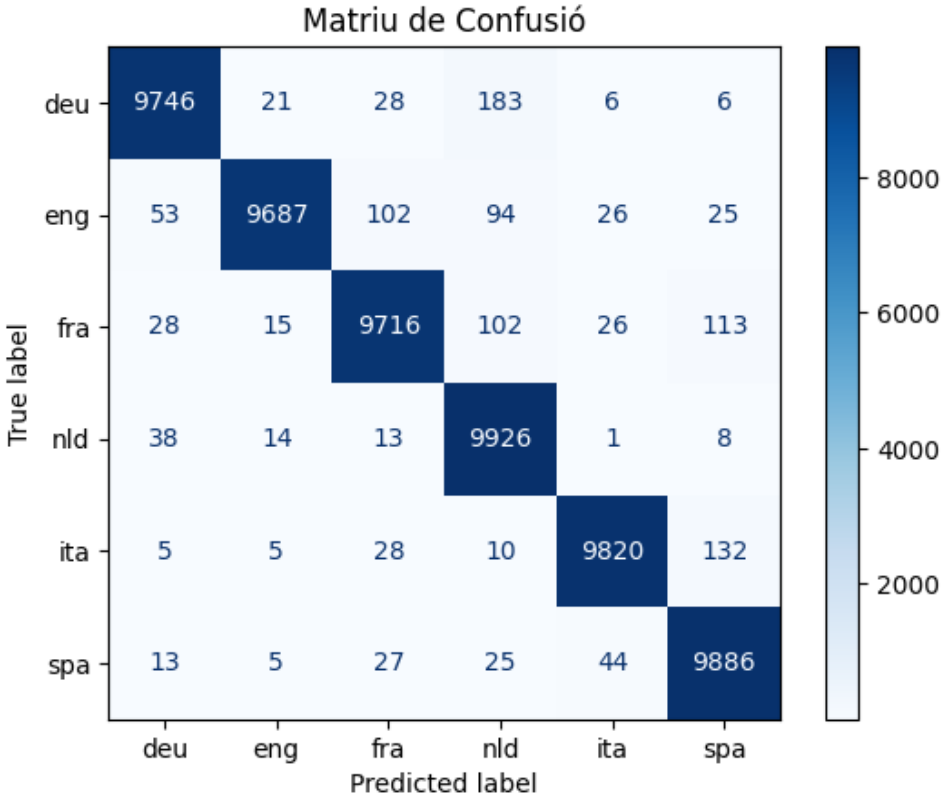

Figure 1: Matriu de confusió del primer experiment.

Com podem observar en la matriu de confusió, la majoria de valors es troben a la diagonal, indicant una precisió significativa en les prediccions. Fent una ullada a la Figura 1, es veu ràpidament que les llengües que millor predeix són el neerlandès, l'italià i l'espanyol. Les altres llengües, com l'alemany,l'anglès i el francès, també mostren una alta precisió, encara que lleugerament inferior.
El model presenta l'error més gran amb 183 instàncies confonent l'alemany amb el neerlandès, indicant una tendència a predir frases amb alemany com si fossin neerlandeses. A més, es detecten conclusions entre l'anglès i el francès, així com entre el francès i el neerlandès i l'espanyol. Tot i així, els resultats globalment són bastant bons.
És important destacar que, malgrat algunes confusions, la major part de les prediccions són correctes, reflectint una eficàcia general del model en la classificació de les llengües.

```
Accuracy: 0.9800590226253397
deu - Recall: 0.9755755755755756, F2-Score: 0.9776698834339828
eng - Recall: 0.9699609492340042, F2-Score: 0.9746453365529731
fra - Recall: 0.9716, F2-Score: 0.9732740313338942
nld - Recall: 0.9926, F2-Score: 0.985895907826778
ita - Recall: 0.982, F2-Score: 0.983514612503255
spa - Recall: 0.9886, F2-Score: 0.9852501494917282

              precision    recall  f1-score   support

         deu       0.99      0.98      0.98      9990
         eng       0.99      0.97      0.98      9987
         fra       0.98      0.97      0.98     10000
         nld       0.99      0.98      0.99     10000
         ita       0.96      0.99      0.98     10000
         spa       0.97      0.99      0.98     10000

    accuracy                           0.98     59977
   macro avg       0.98      0.98      0.98     59977
weighted avg       0.98      0.98      0.98     59977
```

Figure 2: Informe de classificació del primer experiment.

En aquesta segona imatge, s'aprecien diverses mètriques de rendiment que aporten una visió més detallada del comportament del model. En primer lloc, destaquem el "recall" i la "F2 Score" per a cada llengua. És evident que la llengua que millor detecta el model és el neerlandès, mentre que l'anglès mostra el seu pitjor rendiment, confirmant les observacions prèvies extretes de la matriu de confusió. Respecte la precisió, es manté generalment alta per a totes les llengües, amb una lleugera disminució per a l'italià.

5.2 Experiment 2

El següent experiment mostra els resultats finals del nostre model. Com s'ha comentat prèviament, s'han fet alguns canvis per millorar el rendiment del model. Els paràmetres triats finalment han estat: lambda és 0.1, B serà la mitjana i s'ha aplicat l'eliminació de signes de puntuació.

A la Figura 3 s'observa la matriu de confusió que ens mostra les prediccions del model en comparació amb les etiquetes de veritat. Veiem que els valors diagonals són els valors predits de manera correcta. En canvi els altres han estat els que s'han predit de manera incorrecta.

Observem que els valors del mig varien respecte l'anterior però no de manera extrema. Veiem que en els casos de l'alemany, el francès i l'italià, s'han fet millors prediccions (ara s'han predit correctament unes 30 frases en mitjana de més). En canvi en el cas l'anglès, el neerlandès i el castellà, fa una mica de pitjors prediccions. Tot i això continuen siguent molt bones prediccions.

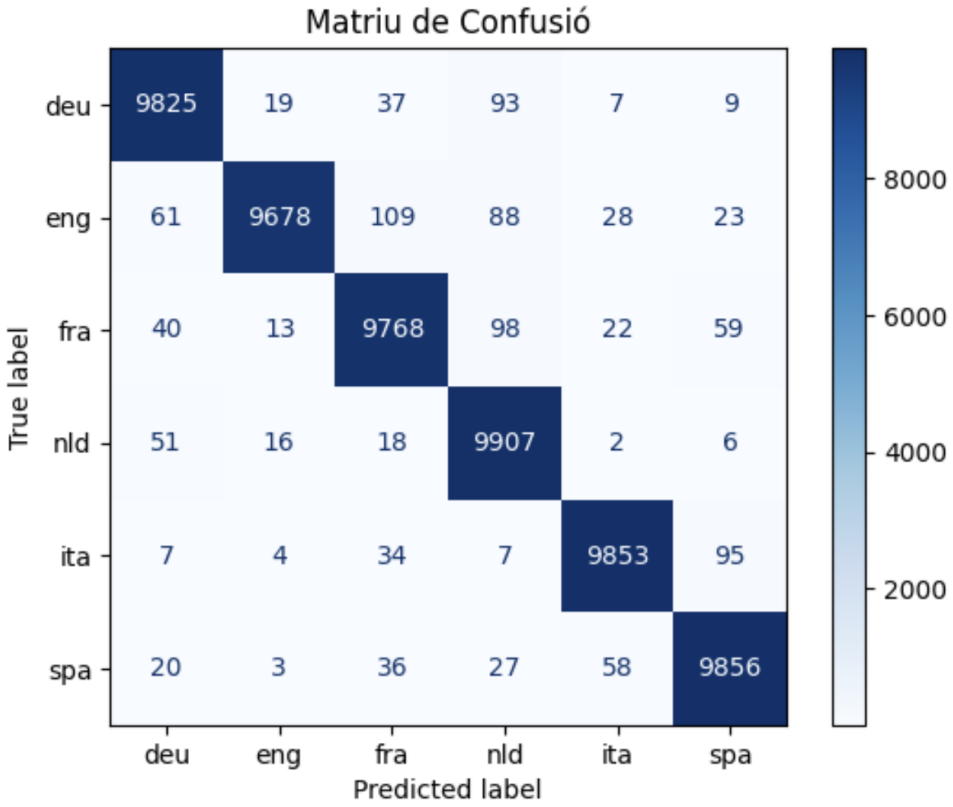

Figure 3: Matriu de confusió del segon experiment.

Veiem que el model no té problemes amb diferenciar entre llengües com l'italià i el castellà amb l'anglès, l'alemany o el neerlandès.

S'observa que hi ha 95 mostres que són predites com a castellà i són italià. També passa entre l'alemany i el neerlandès. I en aquest model veiem que prediu 109 mostres com a francés que en realitat són anglès. Aquest és l'error que fa més gran.

Tot i veure que comet aquests errors, en comparació amb el model anterior, hi ha millores. Per exemple prediu 40 mostres de frases en italià que abans predia a castellà ara ho fa bé. Després passa de predir 183 mostres d'alemany com a neerlandes a la meitat, 93. També és important destacar que abans tenia dificulatats per diferenciar el francés i el castellà i ara en té moltes menys.

A més de la matriu de confusió, també s'han observat diferents mètriques de rendiment (Figura 4).

El model té una accuracy general d'aproximadament 98,18%, que és una millora respecte a l'anterior model que ja era molt bo.

Respecte el recall, veiem que per tots els idiomes està a més del 96%. L'idioma qué té menys recall és l'anglès, i en segona posició el francés. Això vol dir que són els menys capaços d'identificar correctament la majoria d'instàncies certes de cada classe. A l'altre extrem està el neerlandès que té un recall del 0.99%.

La precisió també és molt alta per a tots els idiomes. El que té la precisió més baixa és l'italià. Això vol dir que la majoria de les prediccions d'una classe són correctes.

```
Accuracy: 0.98182636677393
deu - Recall: 0.9834834834834835, F2-Score: 0.9832079096949805
eng - Recall: 0.9690597777110244, F2-Score: 0.9740142106640367
fra - Recall: 0.9768, F2-Score: 0.9767609295628175
nld - Recall: 0.9907, F2-Score: 0.9863600159299083
ita - Recall: 0.9853, F2-Score: 0.9858915349209524
spa - Recall: 0.9856, F2-Score: 0.9846547314578006

              precision    recall  f1-score   support

         deu       0.98      0.98      0.98      9990
         eng       0.99      0.97      0.98      9987
         fra       0.98      0.98      0.98     10000
         nld       0.99      0.99      0.99     10000
         ita       0.97      0.99      0.98     10000
         spa       0.98      0.99      0.98     10000

    accuracy                           0.98     59977
   macro avg       0.98      0.98      0.98     59977
weighted avg       0.98      0.98      0.98     59977
```

Figure 4: Informe de classificació del segon experiment.

Finalment també hem vist oportú analitzar el F1-score i el F2-score. Ens proporcionen una armonia entre la precisió i el recall, i com és alt, sabem que hi ha un bon equilibri entre ambdos.

6 Conclusions

Aquesta pràctica ens ha ajudat extensament a comprendre millor el material teòric presentat a l'assignatura. Ens ha ajudat a veure desde un punt de vista pràctic l'aplicació del que s'ha estudiat. Hem après a fer ús de llibreries molt úlils per al processament del llenguatge.

S'ha fet un preprocessament a les dades per aconseguir un millor entrenament del model. A part del preprocessament indicat a la pràctica, hem afegit l'eliminació de signes de puntuació. L'eliminació de signes de puntuació pot ajudar a simplificar i normalitzar les dades text, ja que pot ser una font de soroll que pot afectar negativament l'aprenentatge del model. Eliminant aquests elements, es pot aconseguir una millor representació de les frases, millorant la capacitat del model per discernir entre les diferents llengües amb més precisió. Aquesta modificació addicional ha demostarat proporcionar millores significatives en la classificació del model.

A l'hora de calcular les probabilitats dels trigrames ha estat clau l'ús d'alguna tècnica de suavitzat per no tenir problemes amb probabilitats 0. En el nostre cas hem vist que tot i que LaPlace no donava mals resultats, utilitzar la generalització, la llei de Lidstone funcionava molt bé. L'ajustament del valor de lambda ha estat crucial per millorar la generalització del model. També hem observat quin valor del paràmetre B, que és una aproximació del total de trigrames diferents, era millor. S'ha vist que utilitzant un valor mitjà dels valors per tots els idiomes donava millors resultats que utilitzant una B diferent per cada idioma, i això pot ser degut a una millor generalització.

Ha estat una sort tenir un balanceig entre totes les diferents classes. Al tenir una quantitat gran i igual per cada idioma, el model ha evaluat i entrenat les dades de manera equilibrada, i no ha estat sesgat. Sinó, hauriem d'haber realitzat alguna tasca de undersampling o oversampling.

Gràcies al preprocessament aplicat a les dades, i un bon entrenament del model, hem obtingut uns resultats molt satisfactoris, amb una precisió molt alta. Com s'ha vist durant l'anàlisi de resultats, el predictor ha aconseguit predir la majoria de frases bé, i els errors que ha tingut han estat causats per semblances entre els idiomes (sobretot entre idiomes que tenen arrels comuns o influències històriques, per exemple les llengües romanes i les germàniques).