UNIVERSITAT POLITÈCNICA DE CATALUNYA

FACULTAT D'INFORMÀTICA DE BARCELONA

Pràctica: Identificació d'idiomes
Autors: Albert Campos Gisbert, César Mejía Rota, Enzo Biasizzo Serra
Data: 04/03/2024

1. Breu explicació de la pràctica

La pràctica consisteix a crear un model que identifiqui l'idioma d'una oració escrita en una llengua europea (anglès, castellà, neerlandès, alemany, italià o francès). Per fer això utilitzem el 'wortschatz leipzig corpora', que consta de unes 30.000 oracions de train i unes 10.000 oracions de test per a cada idioma. Haurem de pre-processar tot els corpora i, en el nostre cas (degut a ser un grup de 3 persones) hem d'implementar un model fet amb trigrames de caràcters i un altre fet amb n-grames de paraules.

2. Preprocessing

El procés de pre-processament dut a terme per la funció 'preprocessing' té com a objectiu aplicar les expressions regulars per preparar un text d'entrada per al seu posterior anàlisi mitjançant una sèrie de transformacions. El procediment dut a terme per la funció es el següent:

S'empra la funció re.sub() per tal d'eliminar tots els caràcters de puntuació, els digits numerics i els espais en blanc generats per les operacions anteriors. Després es torna a fer servir per tal d'eliminar els salts de linea per dos espais en blanc.

Posteriorment, es converteix tot el text a minúscules mitjançant el mètode lower(), la qual cosa homogeneïtza la representació de les paraules i facilita la comparació entre paraules en majúscules i minúscules.

Una vegada realitzades aquestes accions s'eliminen totes les frases que continguin menys de tres caràcters.

3. Creació del model

Pel que fa als models utilitzats, són un model de trigrames de caràcters i un de bigrames de paraules. Tot i que l'n-grama de caràcters venia fixat a l'enunciat de la pràctica, el de paraules no. Ens hem decidit per fer-ho amb bigrames després d'experimentar amb trigrames i adonar-nos de que amb trigrames no acabava de funcionar. Creiem que amb unigrames el rendiment seria inclús millor, tot i que no ho hem provat.

Per crear el model fem, primerament una funció que troba els trigrames de caràcters i una que troba els bigrames de paraules ('X_trigrams_finder'). Aquestes funcions les utilitzarem tant en la creació i assignació de conjunts de trigrames i bigrames a cada idioma com en la funció de suavitzat.

Per assignar i crear els conjunts de n-grames a cada idioma, obrim tots els arxius de train ('_trn'), els preprocessem i els dividim. Guardem els n-grames en un diccionari on les claus son els n-grames i els valors els formen els counts de dits n-grames (filtrem aquells que apareguin menys de 5 vegades en cas dels trigrames de caràcters). Guardem també totes les 6 llengues com a claus de un diccionari superior, on els valors són els diccionaris de n-grames per a cada llengua. Tenim, doncs, 2 diccionaris amb els n-grames de les llengües, un amb els bigrames de paraules i un altre amb els trigrames de caràcters

3.1. Suavitzat amb la Llei de Lidstone

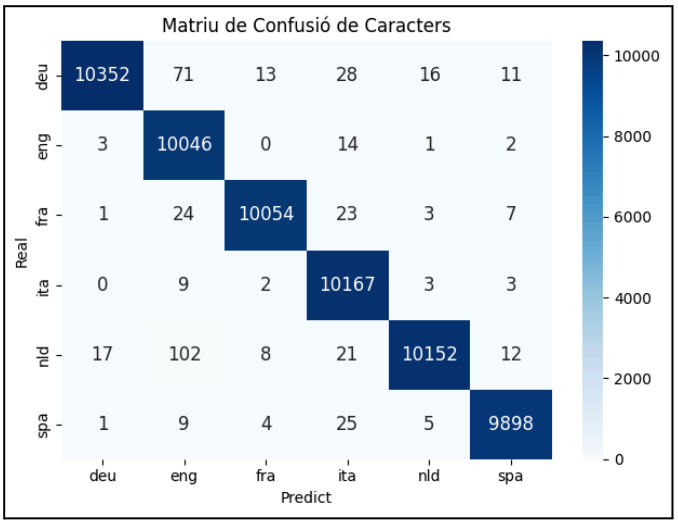
Per a cada n-grama trobat al text entrat, iterem per tots els idiomes buscant el n-grama com a clau del sub-diccionari. Si el troba, utilitza el valor del diccionari per aquella clau (el número de cops que surt el n-grama a les dades del train). Calculem per a cada n-grama el valor del logaritme de la seva funció de Lidstone i fem el sumatori de tots els resultats. Donem com a predicció l'idioma amb el valor més alt. Com a lambda hem agafat una lambda estándar (0.5) y una beta conservadora (la quantitat de n-grames trobats al nostre train). Partim, fent això, de la hipòtesi de que tots els n-grames del llenguatge els coneixem i estan al train. Tot i que això no és cert, evita especular amb la quantitat d'n-grames totals.

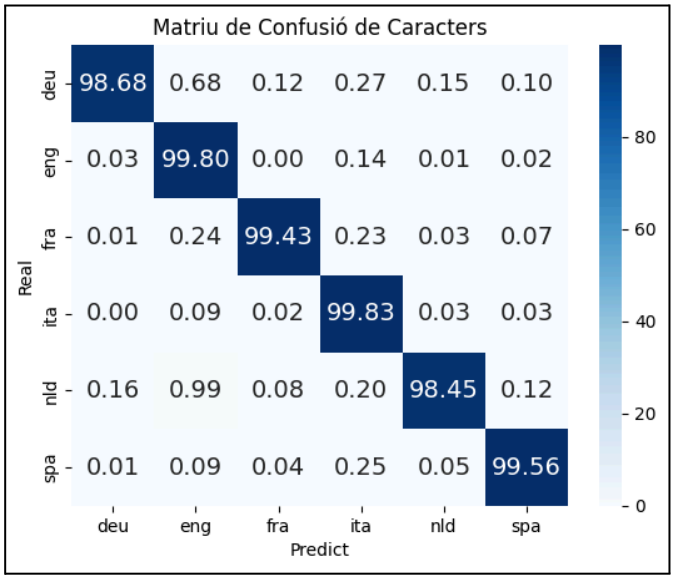
4. El model al test

Per calcular els resultats del nostre model utilitzarem les frases del test. Primer dividim els arxius en frases, després preprocessem cada frase (eliminant les frases amb menys de 3 caràcters) i finalment predim per a cada frase el seu idioma. Guardem els valors reals i els predits i obtenim els següents resultats:

4.1. Resultats del model de trigrames de caràcters

Calculant els resultats obtenim una accuracy del 99,28% i la següent matriu de confusió:

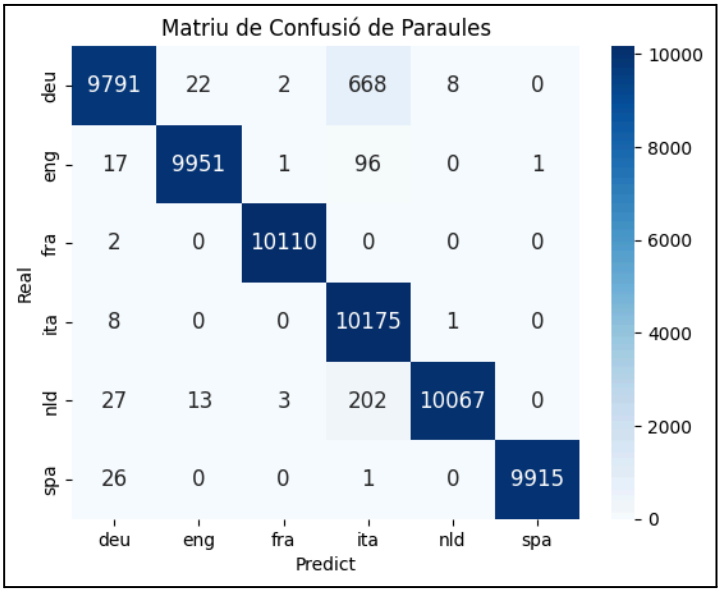




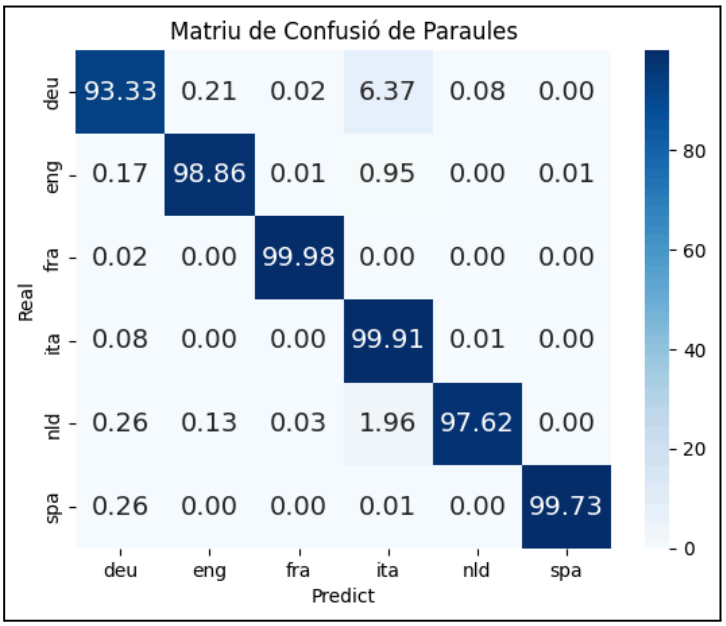
Matriu de confusió amb els valors totals (primera matriu) i matriu amb la precisió per a cada idioma (segona matriu)

4.2. Resultats del model de bigrames de caràcters

Calculant els resultats obtenim una accuracy del 98,20% i la següent matriu de confusió:



Matriu de confusió amb els valors totals obtinguts



Matriu de confusió amb els valors de la precisió per a cada idioma

5. Anàlisi final dels resultats

Els resultats obtinguts en aquesta pràctica han sigut àmpliament satisfactoris.

Per al que fa als trigrames de caràcters, el percentatge d'èxit menor correspon al neerlandés, amb un 98,45%, i en general els valors oscil·len entre 98,45% i el 99,83%, per tant la variació màxima entre llenguatges es al voltant d'un 1%. La precisió total és d'un 99.28%.

En quant als n-grames de paraules, vam començar la pràctica tokenitzant-les per trigrames, però els resultats no eren bons (precisió del 70%), ja que moltes series de paraules que hi eren al test no hi eren als textos corresponents al 'train', per tant, es va optar per utilitzar bigrames de paraules, augmentant així el nombre de coincidències entre els textos del test i el train. Els resultats van millorar considerablement, arribant a un 'accuracy' general del 98,20%.

Creiem que els resultats obtinguts són tan bons perquè els corpora de train i els de test estan altament relacionats a nivell de contingut, tant en vocabulari com en registre i a nivell morfosintactic. Si el test hagués sigut molt diferent al train, tot i que no podem asegurar que no funcionés el model, creiem que el seu rendiment serà pitjor.

Com a conclusió, es considera que el model que funciona mitjançant trigrames de caràcters és més polivalent, i que per tant, podría adecuarse millor en textos que no siguin del mateix àmbit, a més de que ha demostrat una probabilitat d'èxit al volant d'un 1% superior per la tècnica de 'smoothing' aplicada.