STIC-B545 - Traitement automatique de corpus

Etudiant

NGOUFACK NANFAH Angelot

Soumettez sur l'UV un court rapport (2-3 pages, format PDF) présentant brièvement les découvertes relatives à l'année choisie et expliquant ce que le topic modeling pourrait apporter de plus.

Pour réaliser ce traitement on a utiliser pas mal de librairie et d'outils performant :

Extraction des mots clés avec [Yake]

Pour visualiser des mots les plus récurrents [wordcloud]

Pour traiter et extraire les mots vides [nltk]

Pour la reconnaissance d'entités nommées [Spacy]

Sans oublier l'encodage qui risquerait de se planter sans [uni-

code-utf-8] Et l'analyse des sentiments [textblob-FR]

Rapport :

L'année que j'ai choisi est l'année 1970.

Elle contient 100 fichiers, la plage s'étend de 7836 @ 7936, mais pour le code [0 : n+1] d'où [7836:7937].

Et le fichier 7900 correspond au mois du tapis (qui est un mot clé)

Si nous voyons le début du fichier qui correspond au titre , entête du journal on peut lire ceci :

SAMEDI 3 QCTOBRE 1970 OCTOBRE LE MO IS DU TÀ PI S A cette occasion

Remplis de mots clés dont : Les bigames extraites sont entre autres :

['capitaine Mac', 'Mme Lachaud', 'rue Joseph', 'police qu'il', 'jeune homme']

Pour générer un nuage de mot qui fait sens il fallait faire beaucoup d'itération, car les « stop words » initiaux beaucoup nuisaient au résultat de ce fait j'ai ajoute la liste un nombre considérable et d'autre déformation du a l'océrisation :

"les", "plus", "cette", "fait", "faire", "être", "deux", "comme", "dont", "tout",
"ils", "bien", "sans", "peut", "tous", "après", "ainsi", "donc", "cet", "sous",
"celle", "entre", "encore", "toutes", "pendant", "moins", "dire", "cela", "non",
"faut", "trois", "aussi", "dit", "avoir", "doit", "contre", "depuis", "autres",
"van", "het", "autre", "jusqu", "ville","le soir","part","int","dem","app","leurs","terr","min","ecr", "ceux","fem","tel","très ","chez","tél","apr","jeu","vers","déja","prés","but","dès","près","peu","déjà","rien","mèn","bel"]

Les thèmes les plus fréquents de l'année choisie dans le corpus :

[('bruxelles', 1181), ('rue', 1015), ('ans', 650), ('rossel', 627), ('très', 492), ('prix', 462), ('heures', 411), ('soir', 394), ('brux', 394), ('place', 382)]

Tableau de	
C'est la griffe d'un tuniquealent .	36% positive and 0.275% subjective.
Aznavour, on en connaît les qualités et les faiblesses.	neutral and perfectly objective.
Elle-même sera donc détruite.	neutral and perfectly objective.
En dépit de son obsession d'Hérodiade et de Salomé, craignant de s'imiter lui- même, plein de mépris pour la frivolité parisienne, de soh dégoût des * mœurs spéciales » du mécène des Ballets russes et de plusieurs membres de sa troupe, Strauss, dont la prédilection pour la solution de nouveaux problèmes est connue et des difficultés à vaincre, a-t-il rêvé un mimodrame musical ?	21% positive and 2.425% subjective.
Si c'est là un des résultats du prestige qu'a acquis le/concours national V Pro civitate », il faut en souligner le haut mérite.	9% positive and 0.26666% subjective.
La chute s'est encore accentuée cette semaine, le cours ne perdant pas moins die	neutral and perfectly objective.

15 £, en cinq séances, pour s'établir à 430 £.	
La baisse se chiffre ainsi à 43 % en sept ' moisi Le cours est revenu aux aux enviions des niveaux prévalant à fin 1964.	10% negative and 0.15% subjective.
A l'état neutre, les avoirs du Fonds en une monnaie déterminée représentent 75 % du quota du pays concerné.	neutral and perfectly objective.
La presse parle abondamment cette semaine des grands progrès réalésés dans la fusion « Hoogovens- Hoesch », _ plusieurs fois mise en cause précédemment.	30% positive and 0.2% subjective.
L'activité au marché des obligations en eurodollars demeure élevée et la fermeté continue à dominer sous l'impulsion de la baisse des taux officiels, à New York et en Allemagne et de ceux qui sont prévus en Hollande et en France.	neutral and perfectly objective.

Le Topic Modeling est intéressant car elle permet de découvrir les spécificités des différents textes qui constituent un corpus et d'opérer toutes sortes d'analyses et des métriques.

Un modèle pour le moins probabiliste qui permet de définir l'appartenance de documents à des ou thèmes.

Ainsi on peut comprendre qu'un corpus c'est un regroupement de textes, qui sont eux-mêmes une collection de thématiques, et qu'une thématique est un ensemble de mots. Donc selon moi de façon général c'est une façon simpliste de faire une idée d'un corpus et d'apprécier son thématique.

Dans le notebook 9 l'outil (librairie gensim)

Se montre puisant dans la façon il transforme le texte, le modéliser pour dégager des piste d'analyse plus concrètes. Donc on pourrait aller plus loin que le simple nuage de mot qui compte les occurrences, on pourrait faire des comparaisons. Apprécier le style, comprendre certains contextes. Connaitre plus ou moins quels sont les thèmes qui ont été abordés par tel ou tel auteur, chercher à enlever des ambiguïté.

Remarques : Ceux qui me portent à réfléchir.