Participation à DEFT 2019

Jacques Hilbey, Louise Deléger, Xavier Tannier

DEFT, 2 juillet 2019

Tâche 3

Tâche 3 : Âge et Genre

- Des règles !
- spaCy + PyRATA (Hernandez & Hazem, 2018)

- Âge :
  - Déclencheur + nombre (âgé de *, patient de * ans, …)
  - Quadragénaire, quinquagénaire, etc.

- Genre :
  - Homme, patient, Monsieur, testicule, un enfant, prénom masculin…
  - Femme, patiente, Madame, utérus, une enfant, prénom féminin…

- Cohérence entre le nombre d'âges et le nombre de genres

Tâche 3 : Origine

- Des règles !
- spaCy + PyRATA (Hernandez & Hazem, 2018)

- Avec des ordres de priorité :
  - Présenter, souffrir, subir, se plaindre
  - Prise en charge, diagnostic
  - Hospitaliser, admis, consulter pour
  - « pour * <EOS> »

Tâche 3 : Issue

Run 1

- χ^2 sur les N-grammes (N = 1 à 3)
- Sélection manuelle parmi les termes les plus discriminants
- Pour chaque cas :
  - Score pour chaque issue possible
  - Score nul → NUL
  - Égalité → plus fréquent

Run 2

- Sac de mots tf.idf sur la deuxième moitié de chaque cas clinique
- Classifieur multi-classe et paramètres choisis par validation croisée (SVM)

Cas de plusieurs issues : ignoré

Résultats (1/3)

Entraînement

	Précision	Rappel
Âge	0.986	0.953
Genre	0.993	0.983
Issue (*run 1*)	0.693	0.619
Issue (*run 2*)	Exactitude (*Accuracy*)	
	0.524	

Test (résultats officiels)

	Précision	Rappel	F-Mesure	Meilleure F-mesure
Âge	0.980	0.919	0.948	0.948
Genre	0.981	0.974	0.978	0.978
Issue (*run 1*)	0.486	0.405	0.442	0.505
Issue (*run 2*)	0.498	0.492	0.495	

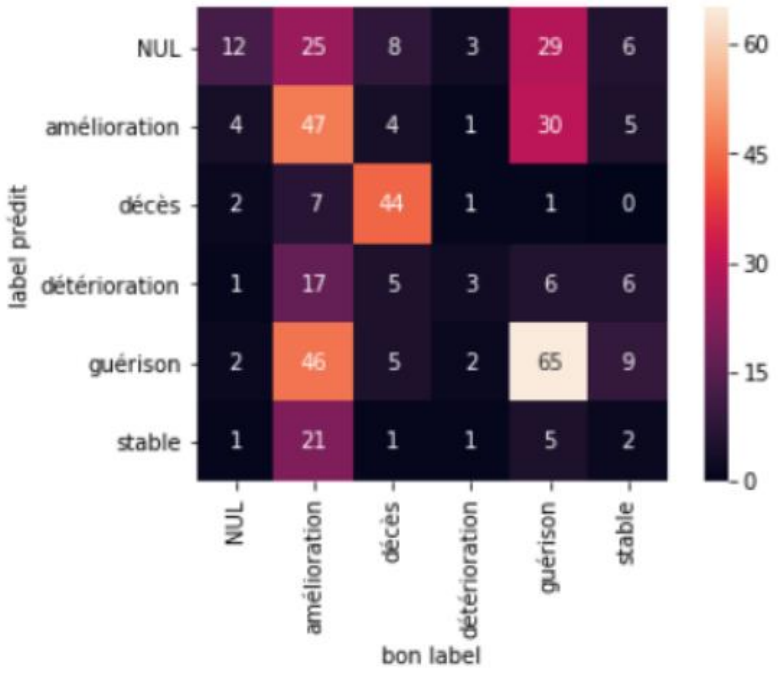
Résultats (2/3)

Test (résultats officiels)

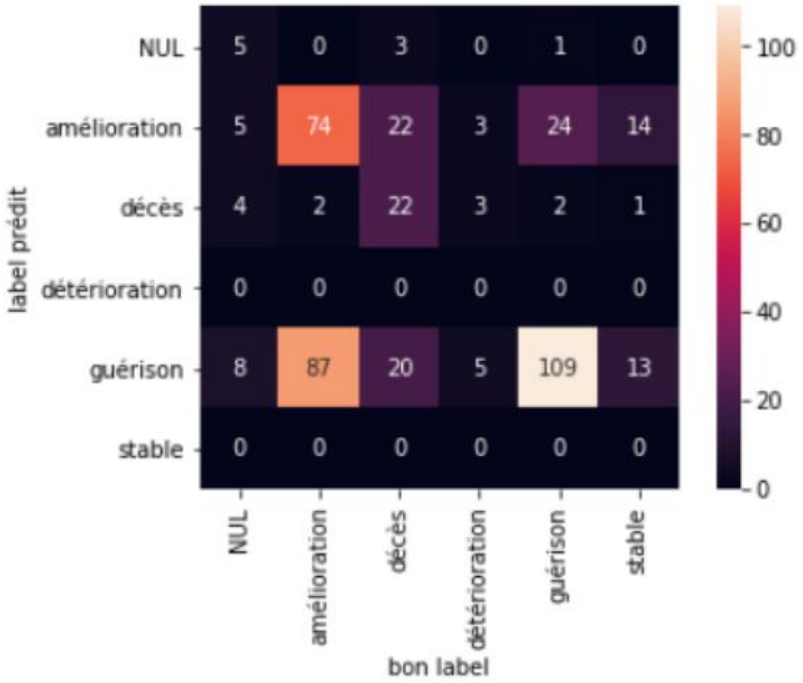
Origine	Macro-	Précision	0.582
		Rappel	0.722
		F-mesure	0.645
		Meilleure F-mesure	0.666
	Micro-	Précision	0.628
		Rappel	0.735
		F-mesure	0.677
		Meilleure F-mesure	0.677
		overlap-accuracy	0.600

Résultats (3/3)

Matrices de confusion pour l'issue



run 1



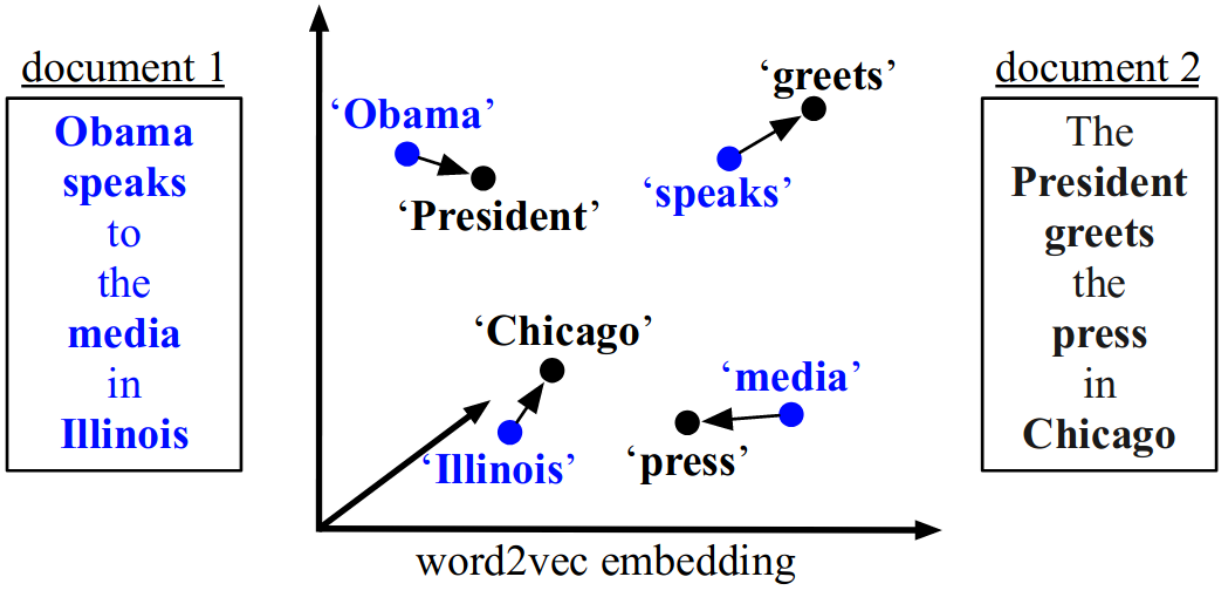
run 2

Tâche 1

(pas de soumission)

Tâche 1

- spaCy + un vilain bricolage pour générer des combinaisons de « GN candidats » de toutes tailles
- KNN avec Word Mover's Distance (Kusner et al, 2015)



Tâche 1

- spaCy + un vilain bricolage pour générer des combinaisons de « GN candidats » de toutes tailles

- KNN avec Word Mover's Distance (Kusner et al, 2015)
  - Embeddings fasttext sur Wikipedia
  - Distance entre chaque candidat du document et chaque terme de la liste
  - Pondération tf.idf pour éviter les termes trop communs de la liste

Tâche 1

- Abandonné car :
  - « meilleurs termes » pas clairs
  - Embeddings pas toujours pertinents :
    - Tumeur →
      - ('tumeurs', 0.7301797468373366),
      - ('lésion', 0.8562563338920136),
      - ('leucémie', 0.8777033952057488),
      - ('infection', 0.8890655330948468),
      - ('inflammation', 0.8989328444873758),
      - ('cancer', 0.9040188121650059),
      - ('pancréas', 0.9066179649613257),
      - ('chirurgicale', 0.9128305713934772)
  - Temps de calcul déraisonnable pour une liste de termes potentiels réaliste