

Projet Traitement Automatique de Texte par l'IA

Peraldi Laurent 22107802

Puura Markus 22001733

Master 1 Informatique

2023/2024



**DIGITAL SYSTEMS
FOR HUMANS**
GRADUATE SCHOOL AND RESEARCH



**UNIVERSITÉ
CÔTE D'AZUR**

Différencier et analyser l'appartenance politique d'un message sur X (anciennement Twitter) dans la politique étasunienne (2017-2018)



Donald Trump
Représentant républicains
des élections 2020



Hillary Clinton
Représentant démocrate
des élections 2020

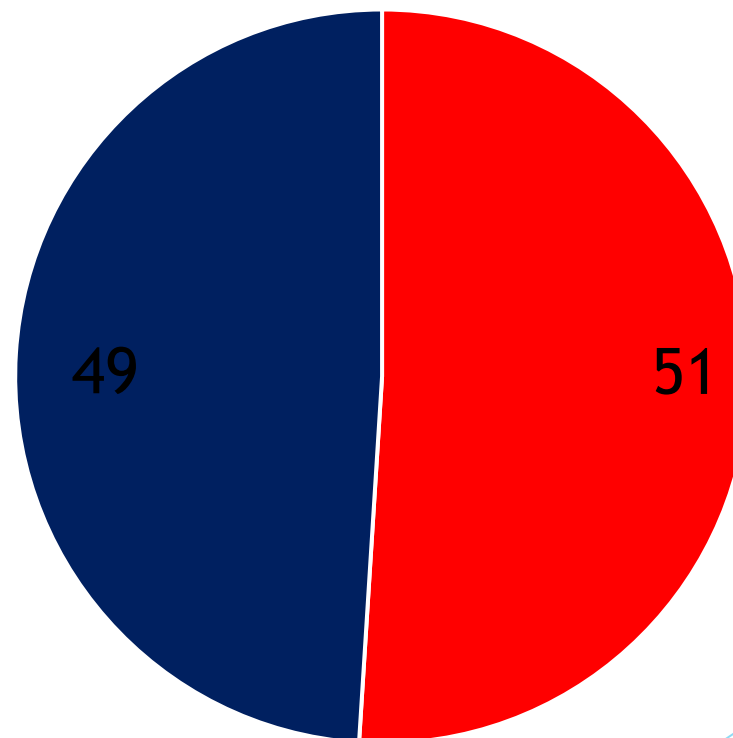
Data set

kaggle

Pourcentage de provenance politique de post X
du data set



85000 post X



Par Kyle Pastor



■ Républicain ■ Démocrate

Tokenisation

Ceci est une phrase d'exemple

↓
Tokenisation
↓

«Ceci» «est» «une» «phrase» «d'» «exemple»

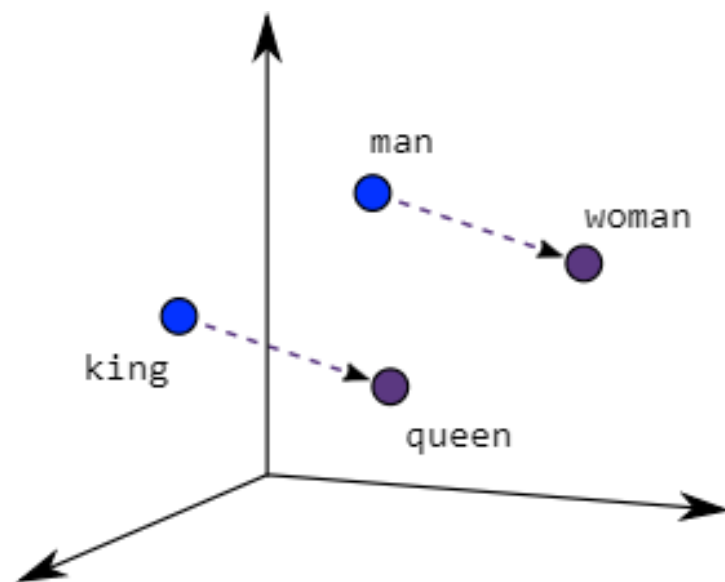
Stemming

«Ceci» «est» «une» «phrase» «d'» «exemple»

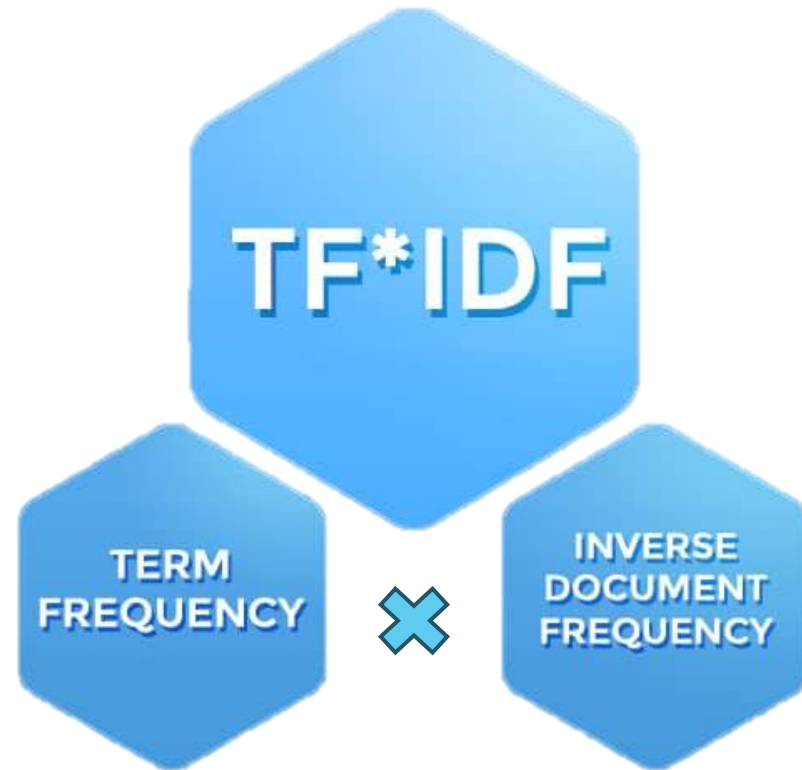
Stemming

«Ceci» «est» «une» «phras» «d'» «exempl»

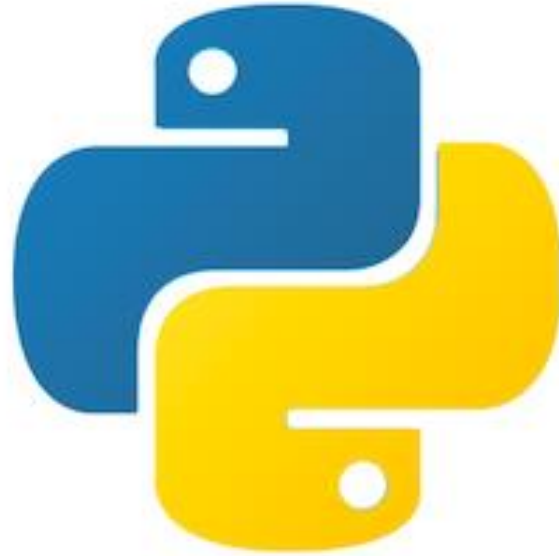
Word embeddings Word2Vec pré-entraîné sur le corpus Brown



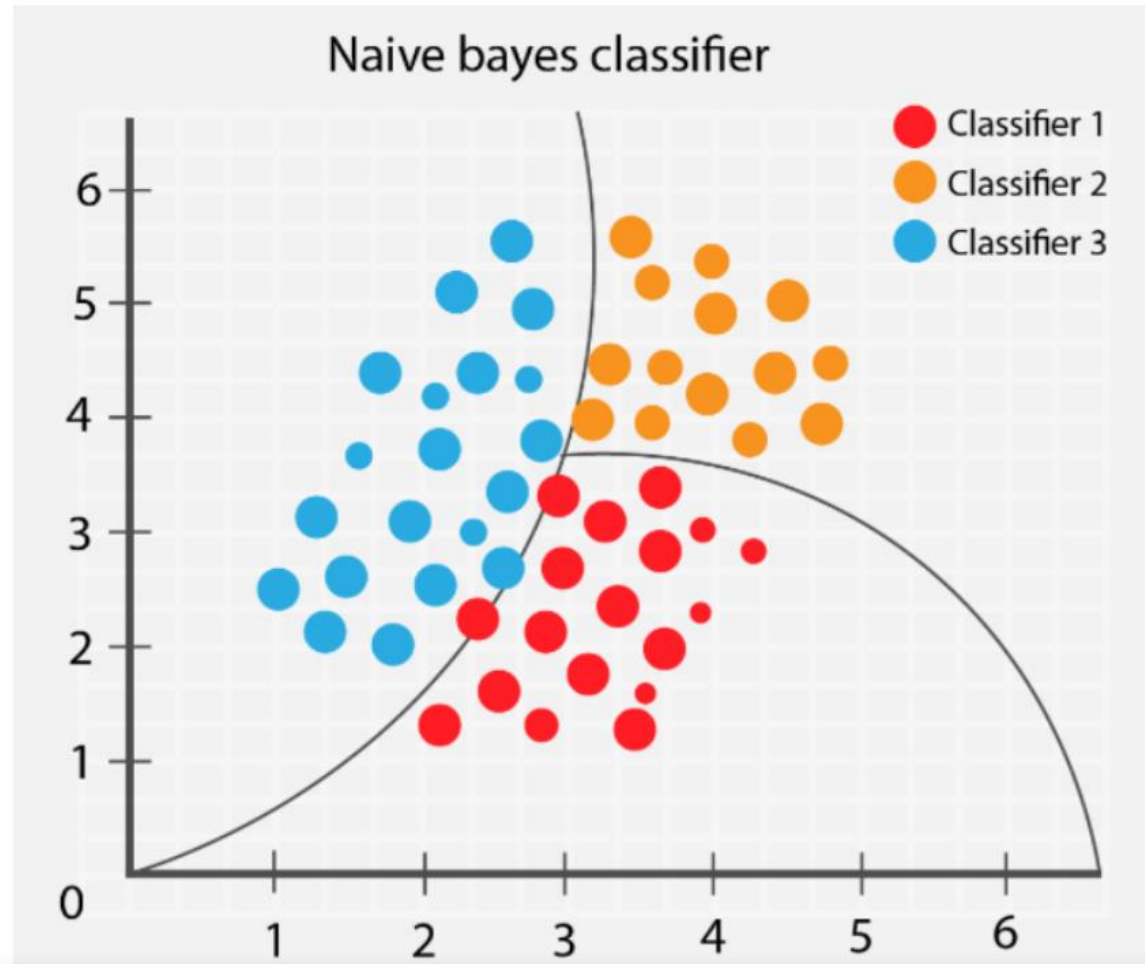
Vectorisation avec tf idf



Entrainement et test avec la librairie scikit-learn



Classification naïve bayésienne



$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Score de précision théorique

- ▶ Classification naïve bayésiennes : 79%
- ▶ Random Forest : 61%
- ▶ Gradient Boosting : 59%
- ▶ LinearSVC : 58%
- ▶ K-NN : 57%

Fonctionnement

Lorem ipsum dolor sit amet, **consectetur** adipiscing elit.

└──────────┘
Républicain

Résultat : Tweet républicain

Praesent condimentum **molestie** suscipit.

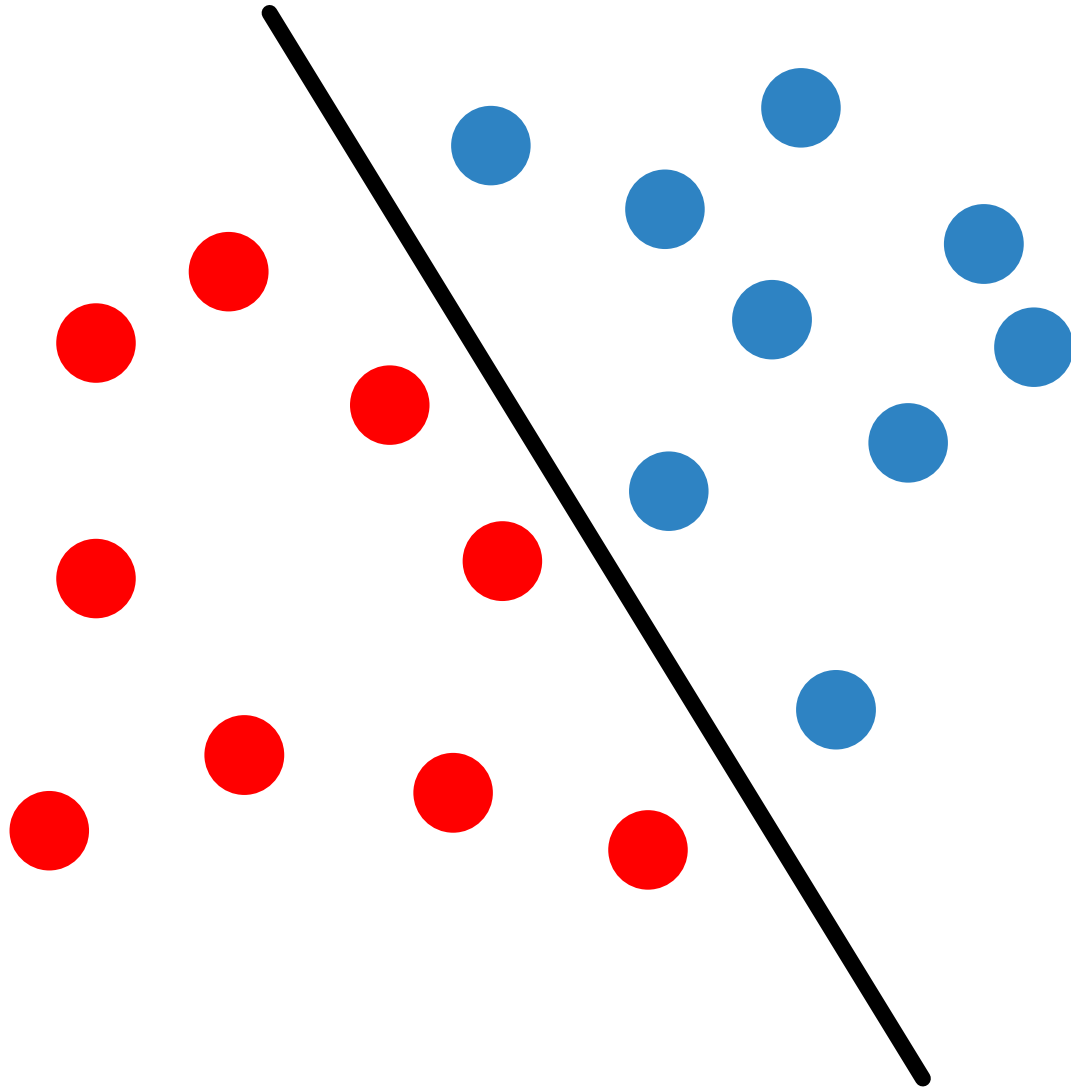
└────────┘
Démocrate

Résultat : Tweet républicain

Quisque ante risus, **faucibus** vitae **scelerisque** egestas, porta non urna.

Dépend du poids du mot

Machines à vecteurs de support linéaire



Support Vector Classification

LinearSVC	Naïve bayésienne
Modèle de ML basé sur des machines à vecteurs de support	Modèle basé sur le théorème de Bayes
Trouve un hyperplan qui maximise la marge entre les classes	
Fonctionne bien dans	

CountVectorizer

Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', ' the', 'lazy', 'dog']



Data

The	quick	brown	fox	jumps	over	lazy	dog
2	1	1	1	1	1	1	1











Résultats - Les mots les plus employés

Top 5 words used by Republicans: ['arkansa' 'bonus' 'fisa' 'lanc' 'schumer']
Top 5 words used by Democrats: ['cbc' 'rm' 'loesack' 'nh' 'pruitt']

Républicain	Démocrate
Dérivé de « Arkansas »	CBC (Cannabichromène)
Bonus	RM (Missed Retweet ?)
FISA (Foreign Intelligence Surveillance Act)	Loesack
Lancaster	NH (New Haven)
Schumer	Pruitt
	Delaware

Résultats - Les @ les plus employés

Top 5 at used by Republicans: ['jim_jordan' 'justinamash' 'housecommmerc' 'arkansa' 'repmarkwalk']
Top 5 at used by Democrats: ['energycommmerc' 'housejuddem' 'cbrangel' 'waysmeanscmt' 'hispaniccaucu']

Républicain		Démocrate	
@Jim_Jordan		@EnergyCommerce	
@JustinAmash		@HouseJuddem	
@HouseCommerce		@Cbrangel	
@ArkansasOnline		@WaysMeansCmte	
@RepMarkwalker		@HispanicCaucus	

Résultats - Les hashtags les plus employés

Top 5 hashtag used by Republicans: ['taxcutsandjobsact' 'taxreform' 'onthisday' 'schumershutdown' 'gopfutur']
Top 5 hashtag used by Democrats: ['goptaxscam' 'ne02' 'netneutr' 'trumpshutdown' 'pruitt']

Républicain	Démocrate
#TaxCutAndJobsAct	#GOPTaxScam
#TaxReform	#NE02
#OnThisDay	#NetNeutrality
#SchumerShutdown	#TrumpShutdown
#GOPFuture	Variation de l'utilisation de hashtag de Scott Pruitt