



DU SDA PROJET NLP

Sylvie Jarjayes

18/12/2022

Objectifs

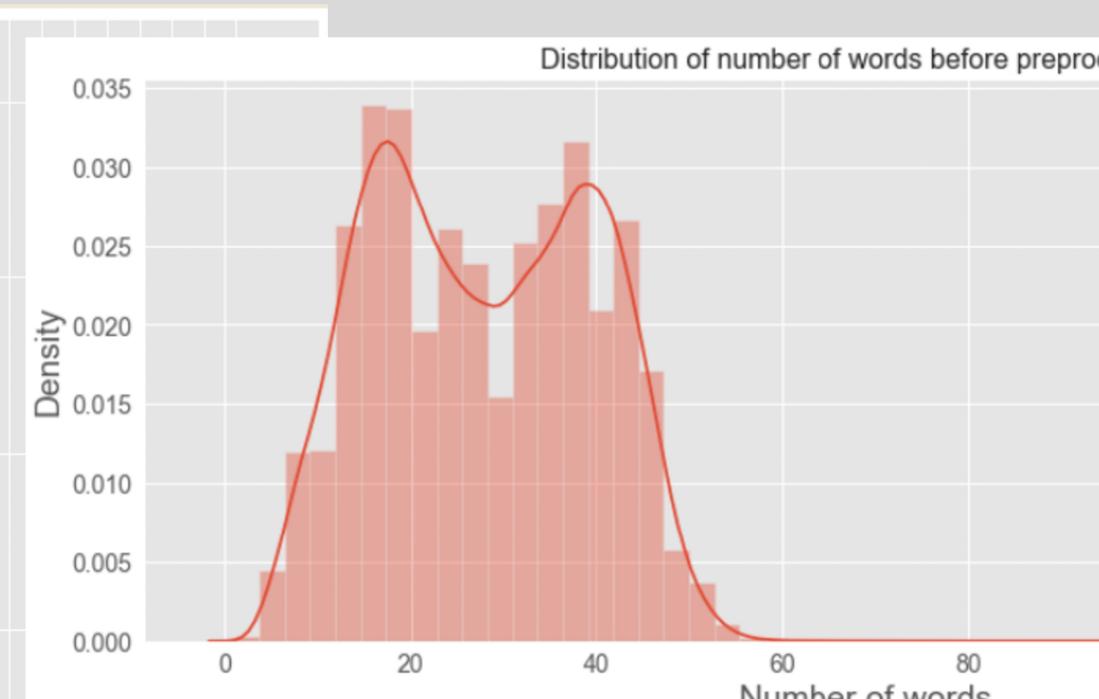
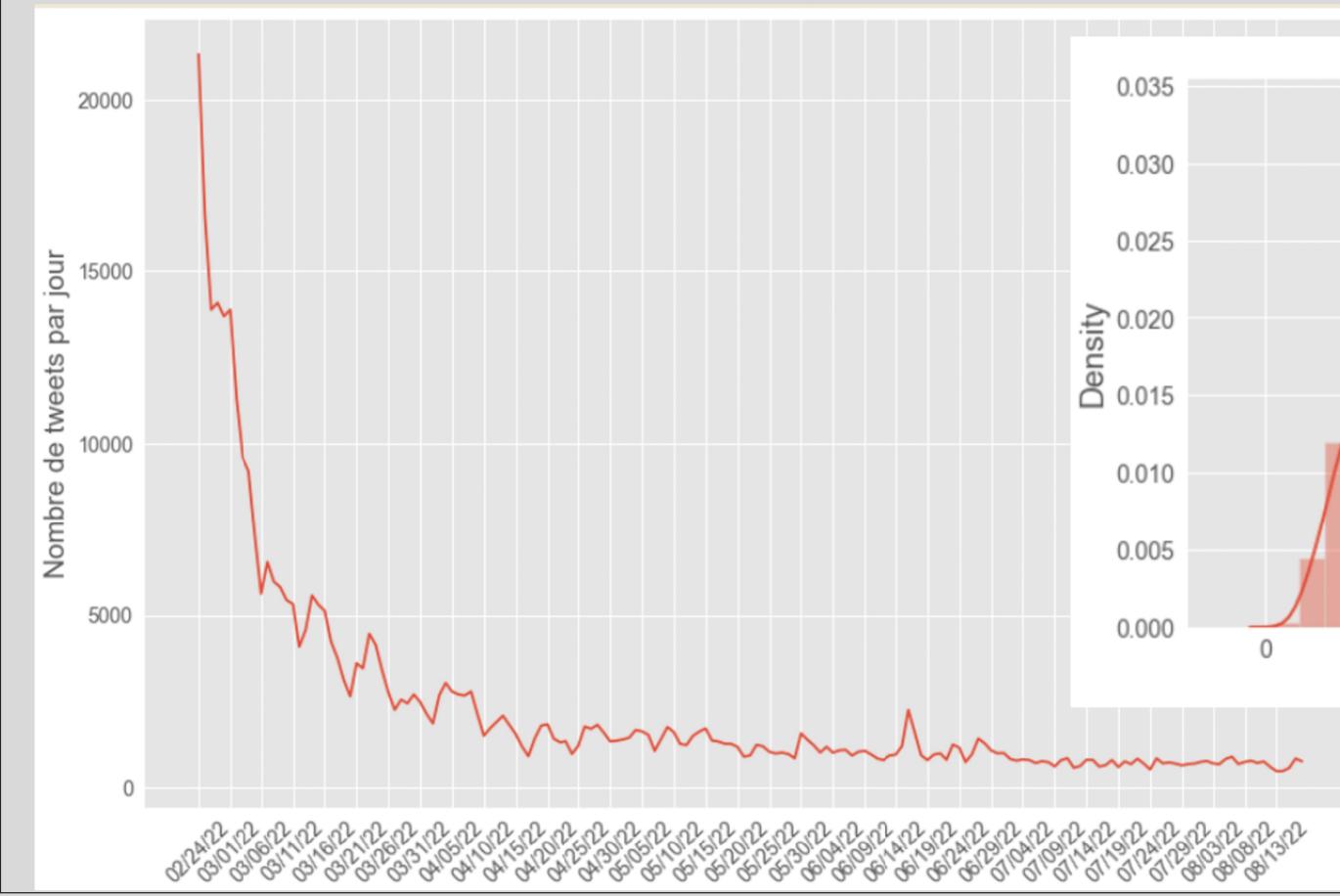
- **Objectif:**
Analyser les tweets en français sur la guerre de 2022 en Ukraine.
 - Importer les tweets avec l'outil adapté : premières analyses
 - Les transformer en données analysables mathématiquement
 - Appliquer des algorithmes de ML à ces données pour mettre en évidence les grandes thématiques :
 - Clustering
 - LDA

Import des tweets avec snscreape: fichier *.json

	_type	url	date	content	renderedContent
0	snscreape.modules.twitter.Tweet	https://twitter.com/M_Degage/status/1560052163212152837	2022-08-17 23:53:14+00:00	#INFO à #RT 🙏 ❤️ \n! #FR #RU #eZ #GJ #JB #Ir \n#Zemmour #Zozz #Patriotes #JamponBeurre \n\n🔴 Jour_175 #GUERRE #Ukraine + #OTAN >> #RUSSIE + #...	🟡 #INFO à #RT 🙏 ❤️ \n! #FR #RU #eZ #GJ #JB #Ir \n#Zemmour #Zozz #Patriotes #JamponBeurre \n\n🔴 Jour_175 #GUERRE #Ukraine + #OTAN >> #RUSSIE + #...
1	snscreape.modules.twitter.Tweet	https://twitter.com/millimagino/status/1560051755286757377	2022-08-17 23:51:37+00:00	#Ukraine / Centrale nucléaire de Zaporijja: Kiev affirme qu'il faut se "préparer à tous les scénarios" https://t.co/PA2qLQQUEb via @BFMTV	#Ukraine / Centrale nucléaire de Zaporijja: Kiev affirme qu'il faut se "préparer à tous les scénarios" bfmtv.com/international/... via @BFMTV
2	snscreape.modules.twitter.Tweet	https://twitter.com/Lejojo66/status/1560049701315018752	2022-08-17 23:43:27+00:00	@WAW_AgainstWar Je me demande ce qu'ils vont dire aux peuples russe ses présentateurs a la co...quand la Poutine va perdre Kherson et la Crimée. En ...	@WAW_AgainstWar Je me demande ce qu'ils vont dire aux peuples russe ses présentateurs a la co...quand la Poutine va perdre Kherson et la Crimée. En ...
3	snscreape.modules.twitter.Tweet	https://twitter.com/IUkrinformEra/status/1560049701315018752	2022-08-17	Guerre en Ukraine : Deux civils tués et sept blessés dans la région de Donetsk \n#\Ukraine	Guerre en Ukraine : Deux civils tués et sept blessés dans la région de Donetsk \n#\Ukraine

Premières analyses

1. Nombre de tweets dans le temps, nombre de mots



Mise en forme des tweets

- Nettoyage du texte
- Suppression des stopwords (nltk+get_stop_words)
- Lemmatisation et stammatisation
- Filtrage des tweets (nbre de mots>4)

cont

- 0 jour gt gt evenements jour intercepte transport munition i
- 1 centrale nucleaire zaporijia kiev affirme faut preparer scenarios vi
- 2 demande dire peuples russe presentateurs co poutine va perdre cas savent bien lecher cul poutine gloire
- 3 civils tués sept blessés region donetsk
- 4 berlin connaissance présence matériel militaire pourrait remis britann americains français voire allemands

content_stem

jour gt gt even jour intercept transport munit i

central nucleair zaporijji kiev affirm faut prepar scenarios vi

demand dir peopl russ present co poutin va perdr kherson crime
cas savent bien lech cul poutin gloir

civil tu sept bless region donetsk

berlin connaiss presenc materiel militair pourr rem britann
americain franc voir allemand

jour gt gt evenements jour intercepter transport munition ie

centrale nucleaire zaporijia kiev affirmer falloir preparer scenarios vi

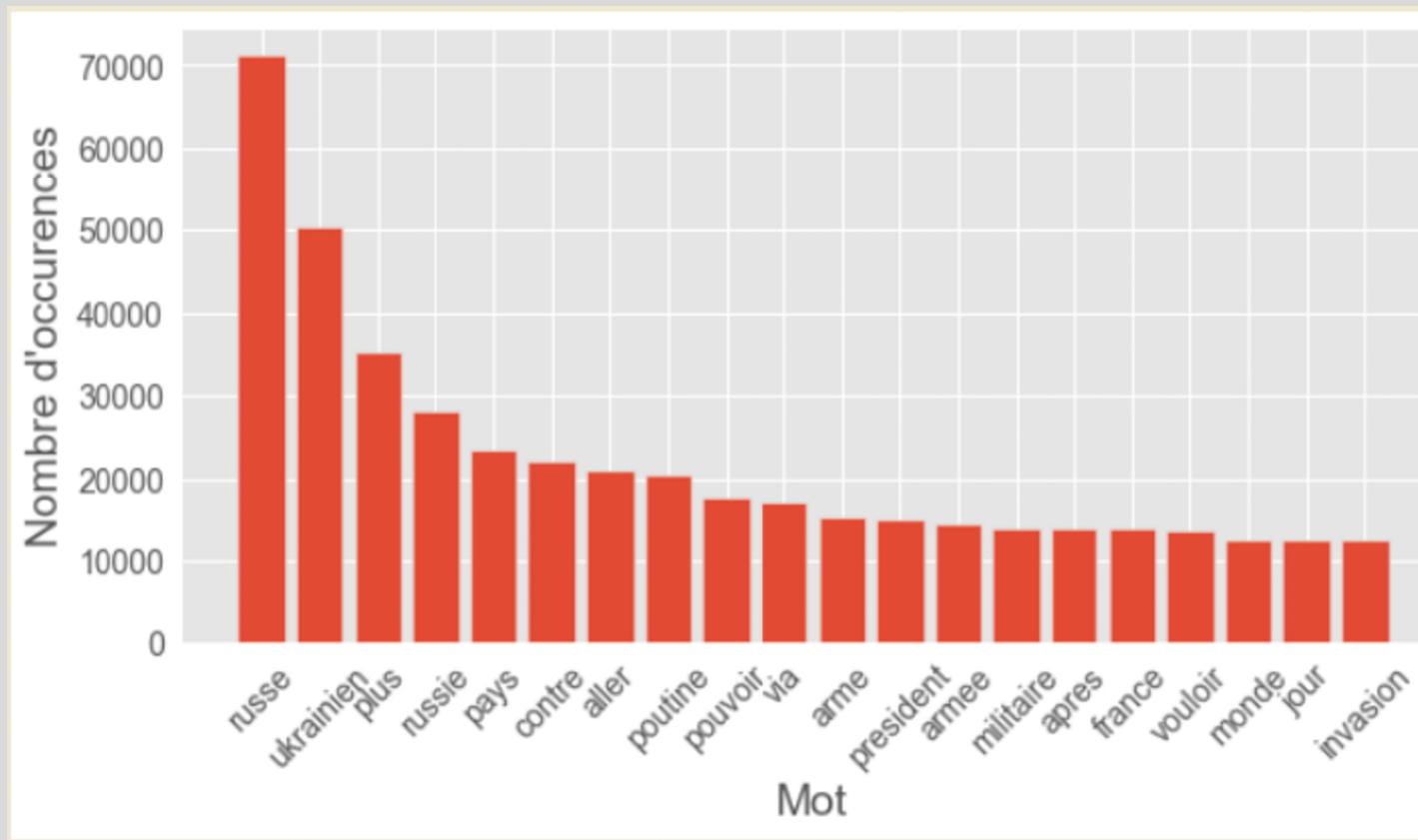
demande dire peuple russe presentateurs co poutine aller perdre kh
cas savoir bien lecher cul poutine gloire

civil tuer sept blesser region donetsk

berlin connaissance présence matériel militaire pouvoir remis britan
americains français voire allemand

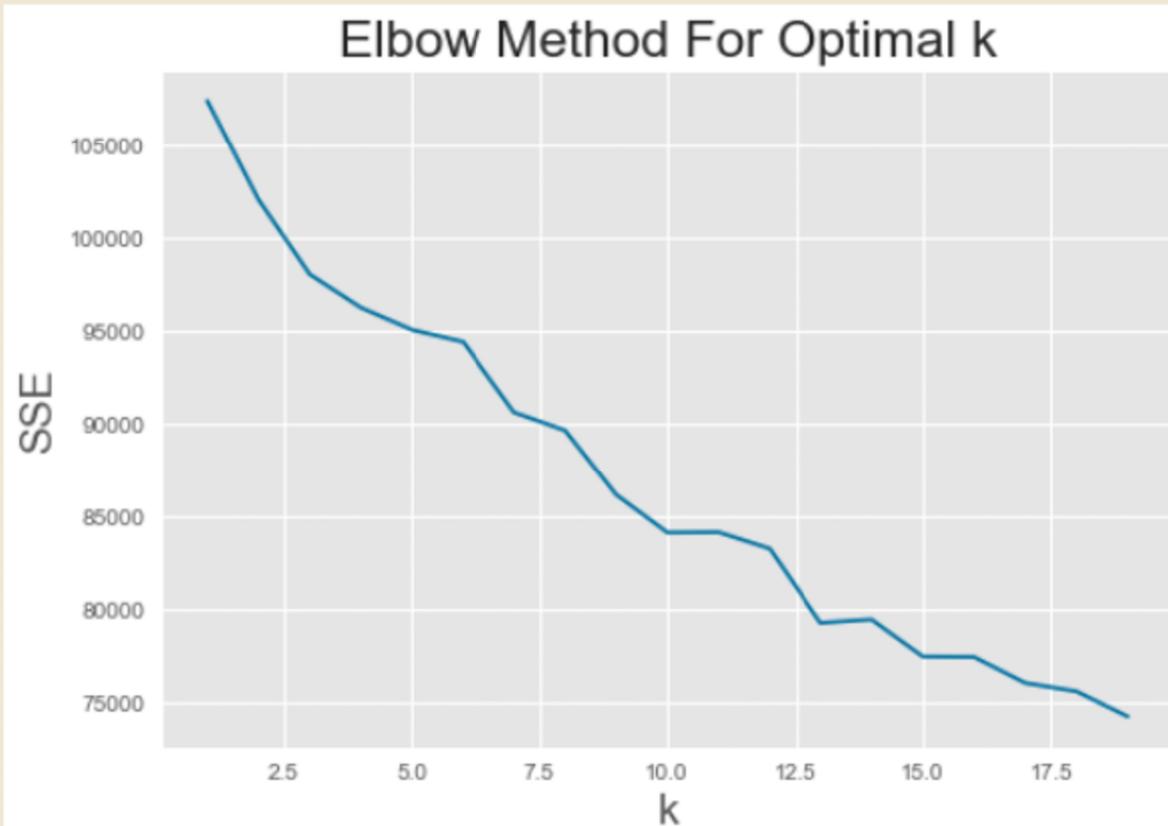
Mots les plus utilisés:

On obtient les mêmes résultats en analysant uniquement les hashtags



Recherche de Clusters:

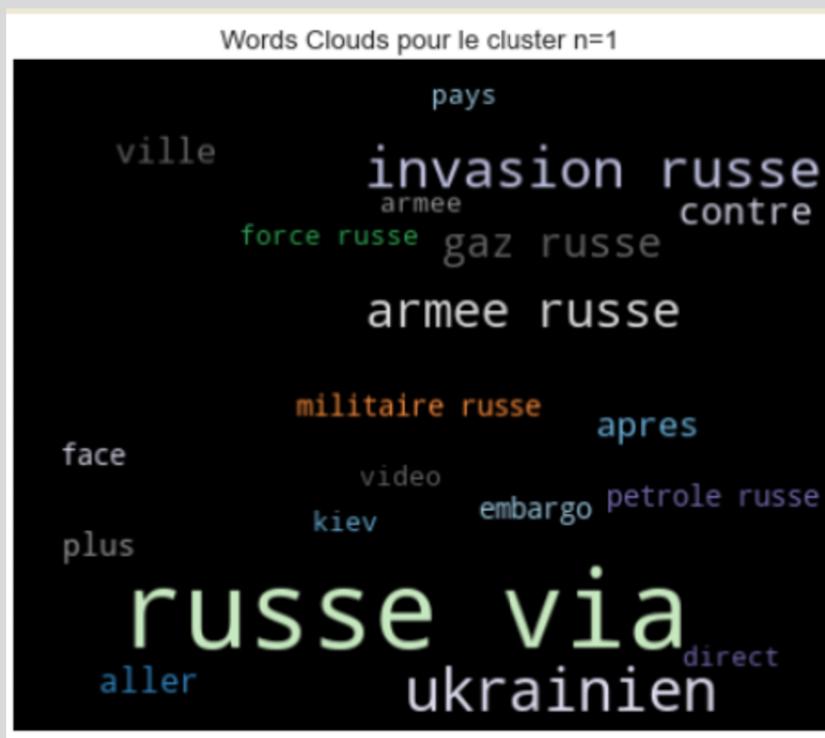
- Vectorisation avec TF-IDF ou Count Vectorizer
- Clustering: Kmeans
- Avec et sans réduction de dimensions



- Changement de pente difficile à
- Etude de l'influence de k

Clusters, sans réduction de dimension:

- Vectorisation
- Clustering: Kmeans
- Avec et sans réduction de dimensions
- Clusters très déséquilibrés



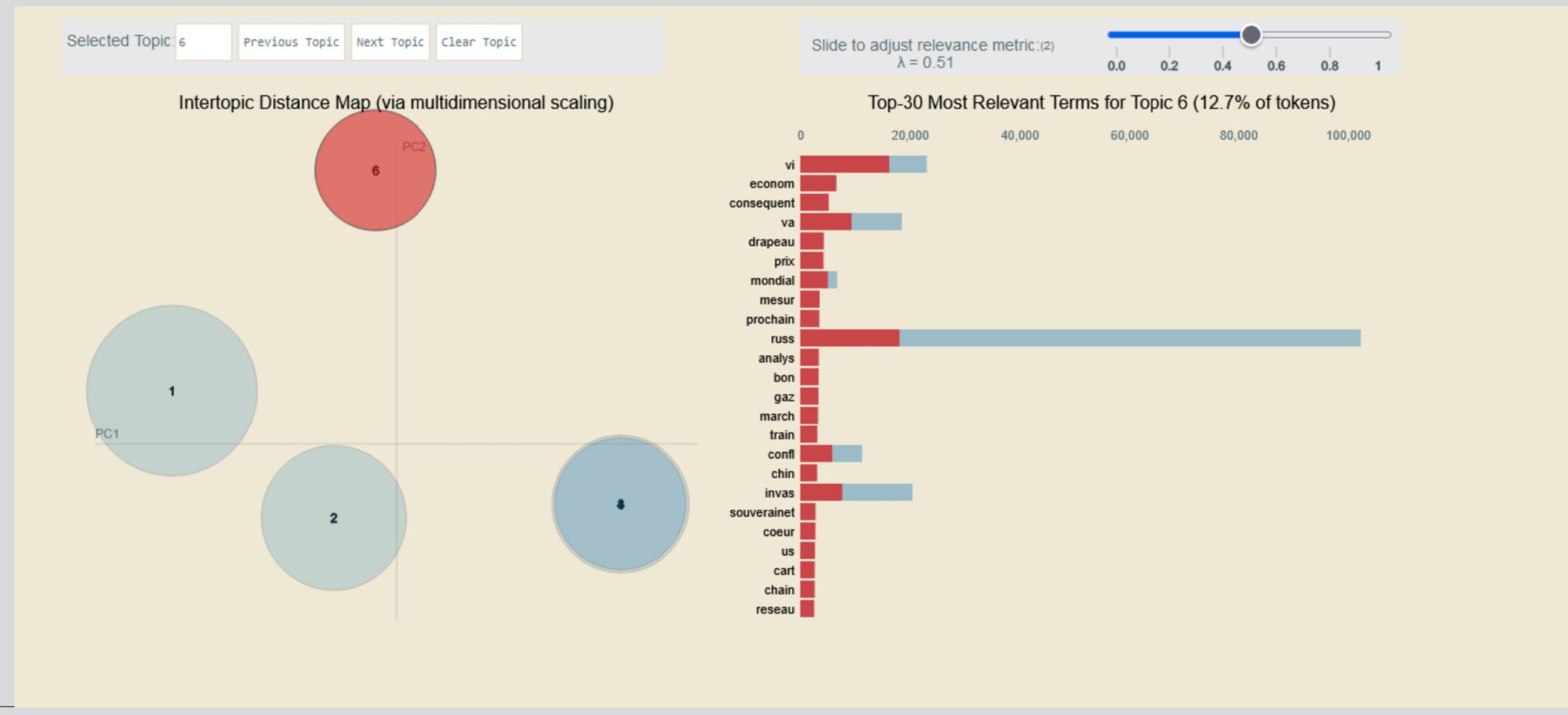
LDA-Analyse et résultats

- Suppression des outliers (mots peu ou trop présents)
- Latent Dirichlet Algorithm, 6 Topics
- Stemmatisation versus lemmatisation

```
[0,  
 '0.046*"russ" + 0.040*"ukrainien" + 0.026*"civil" + 0.025*"bombard" + 0.023*"vill" + 0.020*"direct" + 0.020*"port" + 0.017*"mort" + 0.017*"vi" + 0.017*"kiev"),  
(1,  
 '0.052*"russ" + 0.046*"vi" + 0.027*"va" + 0.022*"invas" + 0.019*"econom" + 0.017*"confl" + 0.016*"europ" + 0.015*"consequent" + 0.015*"plus" + 0.014*"mondial"),  
(2,  
 '0.044*"ukrainien" + 0.035*"pay" + 0.032*"aid" + 0.028*"defens" + 0.024*"aerien" + 0.018*"frontier" + 0.017*"refug" + 0.017*"etat" + 0.014*"don" + 0.014*"franc"),  
(3,  
 '0.023*"plus" + 0.021*"poutin" + 0.018*"mond" + 0.016*"bien" + 0.015*"franc" + 0.013*"va" + 0.012*"faut" + 0.012*"non" + 0.011*"pay" + 0.011*"rien"),  
(4,  
 '0.094*"russ" + 0.056*"arme" + 0.047*"militair" + 0.031*"ukrainien" + 0.024*"open" + 0.022*"h" + 0.021*"attaqu" + 0.021*"kiev" + 0.020*"forc" + 0.015*"troup"),  
(5,  
 '0.047*"russ" + 0.039*"president" + 0.028*"poutin" + 0.027*"contr" + 0.025*"europeen" + 0.024*"ukrainien" + 0.020*"sanction" + 0.018*"vladim" + 0.018*"peopl" + 0.017*"paix")]
```

```
[0,  
 '0.099*"russe" + 0.052*"ukrainien" + 0.036*"attaque" + 0.033*"kiev" + 0.028*"armee" + 0.026*"ville" + 0.025*"militaire" + 0.021*"force" + 0.016*"troupe" + 0.014*"a  
(1,  
 '0.058*"president" + 0.047*"heure" + 0.030*"vladimir" + 0.025*"poutine" + 0.024*"operation" + 0.022*"hui" + 0.022*"aujourd" + 0.021*"annonce" + 0.020*"defense" + 0  
(2,  
 '0.050*"russie" + 0.034*"russe" + 0.033*"contre" + 0.021*"otan" + 0.021*"invasion" + 0.021*"sanction" + 0.020*"europe" + 0.018*"pays" + 0.016*"poutine" + 0.013*"fa  
(3,  
 '0.036*"aller" + 0.025*"plus" + 0.021*"poutine" + 0.019*"bien" + 0.019*"vouloir" + 0.018*"falloir" + 0.018*"dire" + 0.016*"pouvoir" + 0.015*"monde" + 0.014*"rien"  
(4,  
 '0.045*"ukrainien" + 0.038*"russe" + 0.027*"condamner" + 0.024*"homme" + 0.023*"civil" + 0.023*"droit" + 0.022*"mort" + 0.018*"enfant" + 0.017*"contre" + 0.016*"tu  
(5,  
 '0.047*"ukrainien" + 0.037*"soutien" + 0.033*"arme" + 0.027*"peuple" + 0.025*"france" + 0.021*"solidarite" + 0.017*"mondial" + 0.016*"unir" + 0.016*"aide" + 0.015*
```

LDA-Visualisation avec Gensim



Conclusion

- Trouver les bons modules parfois difficile: snscreape(accès avec tweepy), simplemma (incompatibilité spacy),
- Clusterisation et LDA peu efficaces sur l'analyse de tweets ?
- A évaluer: LDA sur cluster, etc...