# IDENTIFICATION OF TOPICS IN PUBLICATIONS WITH NATURAL LANGUAGE PROCESSING

Alhousseynou BALL

alhousseynou.ball@polytechnique.edu

March 19, 2021

# Sommaire

## Sommaire

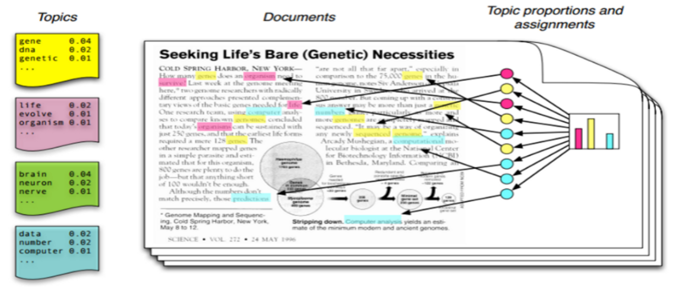
# Identification of topics in publications

- Increase user's commitment rate
- Build new hastags
- +200 000 publications
- Topic modeling: unsupervised learning
- Latent Dirichlet Allocation(LDA) and Non-negative matrix factorization(NMF)

# Identification of topics in publications

## Latent Dirichlet Allocation(LDA)

- Probabilistic model
- Create groups of similar words and find topics

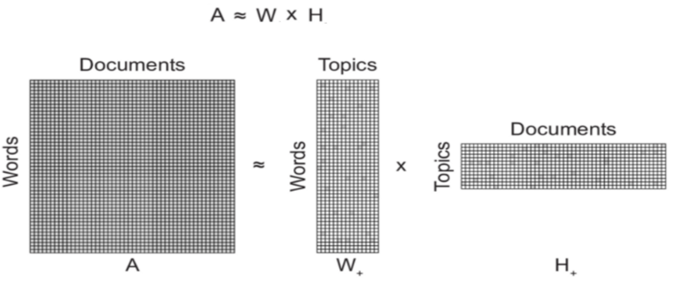



- $\max_{\alpha,\beta} l(\alpha,\beta) = \max_{\alpha,\beta} \Sigma_{m=1}^{M} logp(w_m|\alpha,\beta)$
  - M: nombre of documents
  - $\alpha$ : dirichlet distribution of topics by documents
  - $\beta$ : dirichlet distribution of words by topics
  - w: word

# Identification of topics in publications

## Non-negative matrix factorization(NMF)

- Linear algebra model
- Decomposition of the word-document matrix(A) into two matrices, the first contains all topics and words(W), and the second contains all documents and topics(H).




- $\min_{W,H\geq 0}\{L(A, WH) + P(W, H)\}$ with L is the loss function(Kullback Leibler or Frobenius norm) and P is the penalization(L1 or L2)

# Sommaire

# Data Pre-processing

## Data Pre-processing

- Stopwords
- Outliers: InterQuartile Range method
- Words are lemmatized: words in third person are changed to first person and verbs in past and future tenses are changed into present.

# Sommaire

# Results

## Results: Latent Dirichlet Allocation (LDA)

| Topics | Frequency(%) |
|---|---|
| place-logement-résidence | 29.58 |
| ville-quartier-association | 29.26 |
| fête-loisirs | 22.07 |
| entraide-solidarité | 19.09 |

# Results

## Results: Non-negative matrix factorization(NMF)

| Topics | Frequency(%) |
|---|---:|
| entraide-solidarité | 46.06 |
| horaire-rendez vous | 24.41 |
| fête-loisirs | 8.65 |
| avis-communiquer | 7.46 |
| transport-déplacement | 6.92 |
| qualité de l'air-température | 4.68 |
| entretien-environnement | 1.82 |

# Conclusion

## Conclusion

- Unsupervised learning to identify topics
- Knowing the users' interests
- Build new hastags

# Conclusion

## References


- Alberto Bietti. *Latent Dirichlet Allocation(disponible ici)*, 2012
- Chen Y., Zhang H., Liu R., Ye Z., and Lin J. *Experimental explorations on short text topic mining between LDA and NMF based Schemes.* Knowledge-Based Systems 163, 2019.
- Da Kuang, Jaegul Choo, and Haesun Park. *Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering*, 2014.
- David M. Blei, Andrew Y. Ng, Michael I. Jordan. *Latent Dirichlet Allocation.* Journal of Machine Learning Research 3, 2003.
- Derek Greene. *Topic Modelling With Scikit-learn (disponible ici)*, 2017

# Conclusion

## References


- Lee DD, Seung HS. *Learning the parts of objects by non-negative matrix factorization.* Nature 401:788–791, 1999.
- Paatero P, Tapper U. *Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values.* Environmetrics 5:111–126, 1994.
- Pauca VP, Shahnaz F, Berry MW, Plemmons RJ. *Text mining using non-negative matrix factorizations. In: Proceedings of SIAM international conference on data mining (SDM).* pp 452–456, 2004.
- Rania A, Tet Hin Y and Morad B. *Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis.* Frontiers in Artificial Intelligence, 2020.
- Wikistat. *NMF Factorisation par matrices non négatives(disponible ici).*

# Conclusion

Contacts:

alhousseynou.ball@polytechnique.edu
06 05 63 30 90