*Université Paris Saclay*
*Novembre 2016*

# Centered Title Times Font Size 24 Bold

## Centered Subtitle Times Font Size 16 Bold

**Firstname Lastname**

# Abstract

**Contexte**.
**Objectifs**.
**Méthodes**.
**Résultats**.
**Conclusions**.

**Keywords:** 3-4 keywords, maximum 2 of these from the title, starts 1 line below the abstract.

# Contents

# Chapter 1

# Introduction

# Chapter 2

# Partie 1 - Moteur de Recherche

## 2.1 But

Le but ici était de développer un moteur de recherche capable d'indexer 10 000 documents et de retourner les documents les plus similaires à une requète. Pour cela, des contraines nous sont imposées, le programme ne doit pas utiliser plus de 1GO de mémoire RAM et le fichier index ne doit pas dépasser 60% de la taille du corpus.

## 2.2 Dataset utilisé

Le corpus utilsé ici est celui fournit pour le projet. Un corpus contenant environ 10000 documents datant entre janvier et avril 2015.

## 2.3 Structures de données utilisées

## 2.4 Méthodes utilisées pour l'indexation

L'index contient pour chaque mot, la liste des fichiers auquel il apparait, ainsi que son poid dans ce fichier.

Pour obtnir un fichier index peu volumineux, il a été nécessaire de donner un identifiant aux mots et aux fichiers. Ainsi, nous avons 3 fichiers. Un fichier qui continent les identifiants de chaque mots, un autre qui contient les identifiants pour chaque fichiers, et pour finir l'index contenant pour chaque mot, la liste des fichiers auquel il apparait, ainsi que son poid dans ce fichier.

Là encore, les poids (tfidf) contiennent plusieurs chiffres décimales après la virgule. Pour diminuer la taille de l'index, on a décidé de ne garder que les premier chiffres après la virgule pour chaque poids. Ce choix implique une légère perte de précision.

## 2.5 Méthodes utilisées pour la recherche

## 2.6 Evaluation

### 2.6.1 Pertinences

### 2.6.2 Performances

Pour évaluer, les performances du programmes, l'outil JConsole a été utilisé.

### 2.6.3 Difficultés rencontrés

# Chapter 3

# Partie 2 - Topic Modeling

## 3.1 LDA

## 3.2 Visualisation

## 3.3 Mutual Information

Outline:

1. Définir la problématique: mesurer à quel point le topic modeling est bon
2. Présenter quelques méthodes qui existent pour évaluer les performances du topic modeling
3. Définir Mutual Information/ Pointwise Mutual Information
4. Comment nous allons utiliser la PMI pour
    - Evaluer la cohérence des mots dans un topic
    - Evaluer combien un mot distingue un topic
    - Labeliser les topics avec la PMI
5. Implémentation de la PMI
6. Tests et Résultats

### 3.3.1 brouillon

**Topic Coherence** measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measure- ments help distinguish between topics that are se- mantically interpretable topics and topics that are artifacts of statistical inference [Stevens et al., 2012]

**Mutual information** measures how much information – in the information-theoretic sense – a term contains about the class. If a term's distribution is the same in the class as it is in the collection as a whole, then I (U; C) = 0. MI reaches its maximum value if the term is a perfect indicator for class membership, that is, if the term is present in a document if and only if the document is in the class. [Schütze, 2008]

To apply topic modeling to solve real world problems, we need an easy way to identify "junk" topics (e.g. banana sky canoe furious), which may be statistically well founded, but of no use to end users.

The intuition behind the scoring model comes from the idea that the coherence of a set of ten words implies re- latedness of most or all pairs of words taken from the set. This leads to the idea of a scoring model based on word association between pairs of words, for all word pairs in a topic. But instead of using the collection itself to measure word association — which could reinforce noise or unusual word statistics — we use a large external text data source. Specifically, we measured co-occurrence of word pairs from large external text datasets such as all articles from English Wikipedia,

# Chapter 4

# Conclusion

# References

# References

[Schütze, 2008] Schütze, H. (2008). Introduction to information retrieval. In *Proceedings of the international communication of association for computing machinery conference*.

[Stevens et al., 2012] Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 952–961, Stroudsburg, PA, USA. Association for Computational Linguistics.