**

Projet

*Nuage de mots*

Extraction statistiques sur le contenu de documents.

BARBESANGE Benjamin – GARÇON Benoît

Tuteur : D. HILL — Référent : V. BARRA

15/11/2015

Table des matières

[Introduction 2](#_Toc437262082)

[Nuage de mots 3](#_Toc437262083)

[Présentation 3](#_Toc437262084)

[Applications 3](#_Toc437262085)

[Recherches de l'existant 3](#_Toc437262086)

[Solution présentée 4](#_Toc437262087)

[Les données 4](#_Toc437262088)

[Le programme 4](#_Toc437262089)

[Description 4](#_Toc437262090)

[Fonctionnement 4](#_Toc437262091)

[Le code 4](#_Toc437262092)

[Principe 4](#_Toc437262093)

[La table de hachage 4](#_Toc437262094)

[Outil auxiliaire : génération de phrases 5](#_Toc437262095)

[Intérêt 5](#_Toc437262096)

[Le programme 5](#_Toc437262097)

[Présentation 5](#_Toc437262098)

[Utilisation 5](#_Toc437262099)

[Elaboration du programme 5](#_Toc437262100)

[Principe 5](#_Toc437262101)

[Code 5](#_Toc437262102)

[Conclusion 6](#_Toc437262103)

# Introduction

Ce projet s'inscrit dans le cadre de la seconde année du cycle ingénieur à l'ISIMA. L'intitulé étant l'extraction statistiques dans un document. Par cet intitulé, il y voir le principe de nuage de mots, que l'on peut facilement associer aux mots clés d'un texte. Ceux-ci vont permettre de dégager rapidement la tendance d'un texte.

Cet outil peut avoir de nombreuses utilisations dans le domaine du référencement de documents par exemple ou encore dans la vérification d'écrits scientifiques.

Même si certaines solutions existent déjà, il apparait que certaines ne répondent pas à des exigences comme par exemple la lecture directe de formats de fichiers autre qu'un format de texte brut, ce qui limite les capacités de ces solutions.

# Nuage de mots

## Présentation

Un nuage de mots est une représentation graphique simplifiée d’un texte cohérent. Il regroupe l’ensemble des mots les plus représentatifs du texte en question et organise une hiérarchie de ceux-ci grâce à un jeu de polices et de couleurs. L’extraction des mots-clés peut se faire simplement par analyse statistique du nombre d’occurrences de chaque mot dans le texte ou bien par analyse sémantique.

Historiquement, les nuages de mots sont utilisés à des fins publicitaires et commerciales. Des outils en lignes permettent de générer de véritables visuels professionnels transformant le nuage de mots en un produit commercial. Ils sont aussi utilisés sur le web pour présenter les recherches les plus effectuées sur un moteur pour que l’utilisateur puisse atteindre plus rapidement les résultats les plus populaires.

## Applications

Ce projet peut avoir de nombreuses applications. Certaines peuvent être purement esthétiques pour la présentation simple d'un site web ou d'un document. En effet, le nuage de mot va sortir les mots les plus utilisés, ce qui permet d'avoir la tendance de la resource concernée de manière plus ou moins esthétique en fonction de ce qui est utilisé (ce que nous présentons dans la partie suivante).

Une autre application plus pratique concerne l'indexation des documents. Il est tout à fait possible d'imaginer que lorsque l'on place une resouce en ligne (un rapport de projet, un article scientifique) on souhaite en dégager des mots clés qui serviront ensuite à retrouver plus facilement cette resource.

Enfin une dernière application plus spécifique concernerait les écrits scientifiques. Lorsqu'un article est rédigé par un scientifique sur un domaine précis, il y a certains termes que l'on attend comme dominants. Il dont possible d'extraire les mots les plus utilisés dans cet article et de vérifier si les termes attendus resortent ou non. S'ils ne resortent pas, il est possible d'imaginer que l'article ne concernera pas le sujet et ainsi ne le présentera pas correctement. Ceci peut permettre de gagner du temps lors de la vérification des écrits scientifiques.

## Recherches de l'existant

De nombreux outils existent déjà et permettent de créer des nuages de mots. Après investigations, il est apparu les solutions suivantes.

La plupart des sites se servent d'une zone de texte dans laquelle on colle le texte d'où on veut extraire le contenu. On peut également fournir un lien vers le flux RSS d'un site, ce qui va généralement en donner la tendance.

Le site le plus connu est http://www.wordle.net/. Il permet de créer un nuage de mots à partir d'un texte que l'on colle sur le site. Puis on obtient un résultat graphique, il est important de noter que ce site ne sera bientôt plus compatible puisqu'il repose sur les applet Java qui ne seront bientôt plus supportés par les navigateurs. Un autre site permettant de faire ceci est http://www.tagxedo.com/. Le point faible de ces site et qu'ils reposent sur l'utilisation de plugins et ainsi ne sont pas compatible avec tous les navigateurs.

Certains de ces sites proposent des formes prédéfinies de nuage, et d'autres encore permettent de créer ses prores formes.

Le site http://tagcrowd.com/ possède un avantage par rapport aux autres sites dans la mesure où il est possible d'importer un fichier (au format texte) et d'y extraire le nuage. Voici un exemple crée à partir d'un compte rendu de TP système :



Figure 1 - Nuage de mot à partir de www.tagcrowd.com

Comme on peut le voir, nous retrouvons facilement de thème du compte rendu qui concernait la gestion des droits sur une machine.

Notons que beaucoup de ces outils sont concus pour tirer parti des informations venant de textes en Anglais. C'est pourquoi les pluriels de certains mots Français ne sont pas gérés.

# Solution présentée

## Les données

La présentation d'un nuage de mot ne représentait qu'un bref appercu du mécanisme complexe qui mène à sa réalisation. En effet, le simple fait d'observer les mots les plus courants d'un texte ne suffit pas uniquement à en dégager une tendance. Il faut également prendre en compte d'autres paramètres.

La première chose à laquelle faire attention concerne la langue du texte. En effet, chaque langue va disposer de mots qui sont beaucoup utilisés, comme par exemple en Français : le, la, à, je, etc… Si nous n'ignorons pas ces mots lors de la lecture, nous pouvons être surs qu'ils seront largement dominants dans le retour du programme. Un autre paramètre à prendre en compte va plutôt concernet l'encodage du texte et notamment l'utilisation de caractères spéciaux présents dans certaines langues.

## Le programme

### Description

Nous avons créé un programme permettant l'extraction statistiques des mots d'un texte. Pour ceci, il faut renseigner plusieurs données :

* Le fichier à traiter
* Un fichier contenant des séparateurs (ou utiliser le fichier fourni avec le programme)
* Un fichier contenant des mots à ignorer dans le texte, comme par exemple les auxiliaires et les pronoms

Une fois ces renseignements donnés, le programme va procéder à l'analyse du texte et fournir les résultats selon le mode dont on l'utilise.

Le programme principal peut s'utiliser de différentes manières :

* En mode ligne de commande, ce qui produit un affichage sur une page web
* En mode texte, ce qui permet des interractions directes dans la console et va fournir le résultat à l'écran ou dans un fichier
* En mode graphique, qui l'utilisation aisée du programme et produira des affichages directement dans le fenêtre

### Fonctionnement

## Le code

Pour élaborer cette solution, nous avons dû mettre en place un certain nombre de structures que nous décrirons ici. De plus, d'autre mécanismes ont été introduits et seront également présentés dans cette partie.

### Principe

### La table de hachage

C'est le cœur du programme. Grace à cette structure, nous pouvons trier et stocker facilement les mots puis les compter de manière rapide.

# Outil auxiliaire : génération de phrases

Cette partie du projet sort du cadre initial. Dans cette partie, nous allons analyser le processus menant à la création d'une phrase que nous aurons définie.

Une fois cette phrase définie, le programme va générer aléatoirement une suite de mots qui va ainsi mener à notre phrase.

## Intérêt

L'analyse des différents temps d'obtention de cette phrase va nous permettre de déterminer la part du hasard dans la création de phrases sensées.

## Le programme

### Présentation

### Utilisation

## Elaboration du programme

### Principe

### Code

# Conclusion