**

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.

# 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

## 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.

# Solution présentée

## Les données

## 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