USTHB, Faculté d'Electronique et d'Informatique
Département d'Informatique

Année 2013/2014
Module : Traitement Automatique du Langage Naturel

Projet
Date de Remise: Dimanche 15 Décembre, 2013 à 13h00.

But du Projet :

Le but de ce projet est de vous permettre d'apprécier de façon pratique la puissance des techniques de Traitement Automatique du Langage Naturel combinées aux outils de **Python** de traitement de pages Web.

Avec le développement des Technologies de l'Information et de la Communication (TIC), le problème du plagiat a pris des proportions dangereuses au point où des personnes sans scrupules se sont ises à plagier même des thèses de Doctorat. Certaines ont été rattrapées par le scandale et ont dû payer le juste prix (retrait de diplômes, pénalités financières et même judiciaires, etc.), alors que d'autres sont, pour l'instant « sauves » (mais pourront un jour être rattrapées par le scandale.

Tout comme les TIC permettent de plagier très facilement des travaux d'autrui, ces mêmes TIC permettent aussi de retrouver facilement les preuves du plagiat! Le but de ce projet est de développer un outil anti-plagiat aussi performant que possible. Bien sûr, il existe déjà nombre de ces outils payants ou gratuits, facilement retrouvables sur Internet. Certains sont même Open Source. L'idée est de vous permettre de réfléchir au problème, analyser toutes les formes de plagiat qui peuvent être détectées et que vous pourrez implémenter en utilisant les techniques que vous avez apprises combinées aux outils disponibles dans Python+NLTK.

Spécifications

Votre programme aura les caractéristiques suivantes :

- Une interface agréable (de préférence une fenêtre dédiée aux différentes fonctionnalités)
- Chargement du fichier/document que vous voulez analyser à partir de n'importe quel dossier sur votre ordinateur
- Détection de plagiat à partir de recherche sur Internet. (Plus l'étendue de votre recherche sera grande sur Internet, le mieux ce sera pour votre programme)
- Détection de plagiat à partir d'anciens documents analysés par votre programme. Il faut donc penser à une sauvegarde de tels documents pour détecter les cas de « recyclage » d'anciens travaux ou des travaux d'autres personnes (étudiants, enseignants) qui ont utilisé votre programme (par exemple au niveau de l'USTHB)
- Une définition de votre notion de similarité entre documents ou parties de documents. (Vous pouvez bien sûr vous inspirer de vos lectures de ce que les logiciels anti-plagiat font.)

Il vous est laissé la latitude d'innover dans la solution que vous proposerez, l'innovation étant très appréciée et hautement considérée!!

Ce qui vous est demandé :

Ce projet DOIT être fait par groupes de trois personnes, ce qui veut dire une **pénalisation** des autres formes de coopération. En aucun cas, je n'accepterai plus de 3 personnes sur le même projet. **Vous indiquerez clairement qui a fait exactement quoi sur quelle partie et quelles fonctions de votre programme**. Je n'accepterai pas la mention « nous avons travaillé ensemble sur telle fonctionnalité». Chaque membre de l'équipe prendra entière responsabilité sur certaines parties du programme (conception et implémentation) quand bien même vous les auriez discutées ensemble (bien sûr !). Chacun(e) devra être capable de répondre entièrement aux questions qui tombent sous sa responsabilité.

Il vous est demandé dans ce projet ce qui suit :

1. Commencez par une lecture générale des chapitres pertinents du livre « *Introduction to Natural Language Processing with Python* » et une recherche sur Internet sur les instructions de python qui permettent de créer une fenêtre avec des boutons (qui peuvent être actionnés), etc.

2. Faites une lecture sur la définition du plagiat. Relire le chapitre 0 de ce module et lisez les explications données sur le site www.indiana.edu/~istd/definition.html

3. Faites une petite recherche sur les logiciels de détection de plagiat qui sont disponibles sur Internet pour voir ce qu'ils permettent de faire.

4. Tout votre programme sera **en Python** et permettra:

    a. de lancer une interface qui permettra à l'utilisateur:
        i. de faire son login (userid et password)
        ii. de charger un document et l'analyser du point de vue plagiat par rapport à l'Internet et par rapport à tous les documents analysés précédemment par votre programme (tous les utilisateurs inclus)
        iii. de recevoir un état complet de l'analyse du document qui montre chaque partie problématique et une référence (URL de la page web ou chemin du document) qui montre qu'un plagiat a eu lieu (avec peut-être un taux de similarité)
    b. d'envoyer un email à une personne (par exemple, un enseignant !! ☺) dont l'adresse email peut être mentionnée dans votre logiciel et à laquelle l'utilisateur n'a pas accès. Cet email inclura le document analysé, le nom de l'utilisateur, ainsi que l'état d'analyse anti-plagiat que votre programme a généré.

Ce que vous devez remettre :

5. Chaque équipe doit remettre un rapport et un CD.
    a. Le rapport imprimé qui doit expliquer votre solution : la définition de plagiat que vous avez utilisée, la notion de similarité bien définie, la partie d'Internet couverte par votre recherche (ou bien le corpus que vous avez créé pour simuler cette recherche ainsi que la taille en mots de chaque fichier de ce corpus), une explication des aspects techniques de votre solution (recherche, optimisations, etc.), une évaluation de la précision de votre programme, etc.
    b. Un « manuel d'utilisation » en 2 pages maximum, ajouté en Annexe du rapport, expliquant pas par pas comment faire fonctionner les différentes parties de votre système.
    c. Une mention claire et précise dans le rapport de qui a fait quoi (ces tâches doivent être nettement distribuées et chacun(e) sera questionné lors de la démo sur sa partie).
    d. Une copie électronique du rapport.
    e. Tout le code Python dans un dossier à part nommé « Code Python ». Les noms des dossiers doivent être clairs (l'application, les documents déjà analysés, les résultats de l'analyse, etc.) selon votre solution.
    f. La « déclaration de non-plagiat » dûment remplie et signée individuellement (c.-à-d. par chaque membre de l'équipe) attestant que votre rapport ainsi que votre code est entièrement votre travail sauf là où des références explicites et exhaustives sont données. (Le formulaire de cette déclaration vous sera envoyé prochainement, avant l'échéance de remise de votre projet)

Date de remise du projet (rapport + CD) : **Dimanche 15 Décembre, 2013.**

N.B. :

1. Je n'accepterai aucune remise du projet ou parties du projet par email ; vous devrez tout remettre dans ma boite postale au niveau du département.
2. Pour chaque jour de retard, vous perdrez 25% de votre note.
3. . **Aucun projet ne sera accepté après le Jeudi 19/12/2013 à 13h00.**
4. **J'utiliserai l'outil anti-plagiat MOSS (**http://theory.stanford.edu/~aiken/moss/**) pour vérifier que le code que vous remettrez est bien le vôtre !!! Aucune copie de code de l'Internet, de vos camarades de section, ou autres ne sera tolérée.**
5. La note de votre travail prendra en considération la qualité du travail (analyse, programmation, modularité, commentaires, choix des noms de variables et fonctions, etc.), l'interface de l'outil, les fonctionnalités, la richesse des options disponibles, etc. La note sera divisée comme suit :
    a. Votre rapport comptera pour 50% de la note finale ; le rapport ne doit pas dépasser 10 pages et doit être bien présenté, très lisible, très clair, présentant l'analyse, la conception, la structure des pages web, les algorithmes, la structure de la base de données, etc. En Annexe vous ajouterez tout le code imprimé.
    b. Vous serez appelé à faire une démo de votre système ; cette démo comptera pour 50% de la note. Vous aurez le droit de continuer à améliorer votre système jusqu'au jour de la démo auquel cas vous remettrez un nouveau CD si vous avez modifié votre implémentation. (La note du rapport ne sera pas changée avec ces modifications ; seulement la note de la démo en sera affectée.)
6. La note de chaque membre de l'équipe ne sera pas nécessairement la même, chacun(e) selon sa contribution et sa compréhension du problème, de la solution, etc. De plus, votre note finale sera relative à celles des autres étudiants de la section et sera calculée en triant toutes les notes préliminaires. Ceci veut dire que vous devrez considérer les autres équipes comme vos concurrentes.

Bonus :

- Toute innovation intéressante sera bien récompensée.
- Les efforts qui montrent un traitement de texte pertinent seront généreusement récompensés.

A vous de faire une bonne analyse du problème.