 **Université Mohammed V de Rabat **

**Ecole Nationale Supérieure d’Informatique et d’Analyse des Système**

Rapport du projet fin d’année

**Développement de micro-services Advisor pour les équipes de recherche**

# Réalisé par : Encadré par :

ABOUNACEUR Younes Mme EL ASRI Bouchra

AIT-SSI Mouslim

1

REMERICIEMENTS

Nous tenons à remercier sincèrement :

D’abord Mr EL Asri Bouchra, professeur à l’ENSIAS, pour consacrer son temps et ses efforts pour nous encadrer dans ce projet et de l’examiner. Nous remercions également tous les professeurs qui nous ont enseigné durant ces deux semestres, pour le savoir qu’ils nous ont transmis et pour leur entière disponibilité.

Sans oublier de remercier infiniment le directeur Mr ESSAIDI Mohammed, et tout le corps administratif pour leur dévouement au corps estudiantin.

En espérant que ce travail sera à la hauteur de vos attentes.

Cordialement Younes & Mouslim.

2

Resume

Le sujet de notre projet de fin d’année consiste à développer un système de recommandation qui recommande aux chercheurs les conférences adéquates pour un travail de recherche donné.

Durant ces derniers mois, grâce aux efforts de notre encadrante Mme El Asri, nous avons appris les principes fondamentaux des systèmes de recommandation, Machine learning et la modélisation des données.

Nous vous présenterons donc tout au long du rapport, les étapes de la réalisation de ce projet, ainsi que les outils que nous avons suivi pour réaliser notre projet de fin d’année.

3

ABSTRACT

The subject of our year-end project is to develop a recommendation system that recommends to researchers the appropriate conferences for a given research work.

In recent months, thanks to the efforts of our supervisor Ms. ElAsri, we have learned the fundamentals of recommendation systems, machine learning and data modeling.

We will therefore present you throughout the report, the steps we have followed and the tools we have used to carry out our end-of-year project.

4

SOMMAIRE:

Chapitre 1 :

1.1-L’intelligence artificielle………………………………………………………….….…7

1.2-L’apprentissage automatique………………………………………………….……7

1.3-Systèmes de recommandation……………………………………………….….….8

Chapitre 2 :

2.1-Scraping……………………………………………………………………………………….10

2.2-Data-Set…………………………………………………………………………………….…10

2.2.1-Fichiers CSV……………………………………………………………….…10

2.2.2-Modélisation des données……………………………………………11

2.3-Content-based recommender systems…………………………………….……11

2.4-Pourquoi le content-based recommender systems ?.......................12

2.5-Les concepts utilisés dans les systèmes de recommandation ………12

2.5.1-TF-IDF…………………………………………………………………………..12

2.5.2-Cosine similarity…………………………………………………………...13

Chpaitre 3 :

3.1-le langage de programmation…………………………………….……………......14

3.1.1-Numpy………………………………………………………………………….15

3.1.2-BeautifulSoup…………………………………………………………….…15

3.1.3-Pandas………………………………………………………………………….15

3.1.4-Sickit Learn…………………………………………………………...........15

3.2-Environements du développement……………………………………………….16

3.2.1-Spyder………………………………………………………………………….16

3.2.2-Sublime Text………………………………………..……………….........16

3.3-Réalisation……………………………………………………………..………………………17

Conclusion…………………………………………………………………………………………………..…....18

Bibliographie……………………………………………………………………………………………………..19

5

introduction generale

Afin d’appliquer les méthodologies et les notions enseignées en 1 ère année de notre cursus de formation ingénieur en informatique à l’ENSIAS, nous sommes invités à réaliser un projet pratique (PFA), ce qui nous permet d’appliquer nous connaissances théoriques sur le champs pratique, et de s’habituer au travail en groupe sous l’encadrement d’un professeur.

Les systèmes de recommandation ont connu un succès énorme durant ces dernières années et ils sont notamment utilisés par les plus grandes entreprises du monde comme : Facebook, YouTube, Netflix et Amazon.

Notre PFA se situe dans ce contexte, il s’agit d’un système de recommandation qui aide le chercheur à identifier les conférences où son article a plus de chance d’être accepté.

En premier lieu nous vous présenterons l’idée générale, le principe et notre approche. Ensuite, nous mentionnerons les outils qu’on a utilisé et nous détaillerons notre code, afin de vous donner une image claire sur le projet.

6

CHAPITRE 1 : contexte génerale du projeT

Ce chapitre sera consacré pour introduire quelques notions fondamentales sur les quelles notre projet est basé.

# 1.1-l’Intelligence artificielle

Comme son nom l'indique, l'intelligence artificielle peut être interprétée vaguement comme une incorporation de l'intelligence humaine aux machines.

L'intelligence artificielle est le concept plus large qui englobe tout, depuis l’intelligence artificielle traditionnelle jusqu’aux technologies futuristes telles que l'apprentissage en profondeur (Deep Learning) .

Chaque fois qu'une machine exécute des tâches sur la base d'un ensemble de règles stipulées qui résolvent des problèmes (algorithmes), un tel comportement « intelligent » s'appelle l'intelligence artificielle.

Par exemple, de telles machines peuvent déplacer et manipuler des objets, reconnaître si quelqu'un a levé la main ou résoudre d'autres problèmes.

Les machines à moteur IA sont généralement classées en deux groupes - général et étroit. Les machines d'intelligence artificielle générale peuvent résoudre intelligemment des problèmes, comme ceux mentionnés ci-dessus.

Les machines à intelligence étroite peuvent effectuer des tâches spécifiques, parfois meilleures que les humains, même si leur portée est limitée.

# 1.2-l’apprentissage automatique (machine learning)

Comme son nom l’indique, l’apprentissage automatique est l’ensemble des algorithmes qui donnent aux systèmes informatiques la capacité d’apprendre.

L'intention de ML est de permettre aux machines d'apprendre par elles-mêmes en utilisant les données fournies et de faire des prédictions précises.

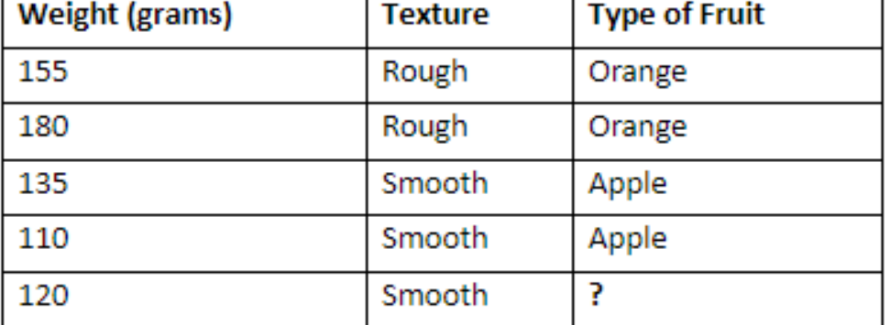
ML est un sous-ensemble de l'intelligence artificielle ; En fait, c’est simplement une technique pour réaliser l’IA.

C'est une méthode de formation d'algorithmes tels qu'ils peuvent apprendre à prendre des décisions.

7

L'apprentissage en machine implique de donner beaucoup de données à l'algorithme et de lui permettre d'en apprendre plus sur les informations traitées.

Par exemple, voici un tableau qui identifie le type de fruit en fonction de ses caractéristiques :



Comme vous pouvez le voir sur le tableau ci-dessus, les fruits sont différenciés en fonction de leur poids et de leur texture.

Cependant, la dernière rangée ne donne que le poids et la texture, sans le type de fruit.

Et, un algorithme d'apprentissage automatique peut être développé pour essayer d'identifier si le fruit est une orange ou une pomme.

Une fois que l'algorithme est alimenté avec les données d'apprentissage, il va apprendre les différentes caractéristiques entre une orange et une pomme.

Par conséquent, s'il dispose de données de poids et de texture, il peut prédire avec précision le type de fruit présentant ces caractéristiques.

# 1.3-systèmes de recommandation

Les systèmes de recommandation font partie des applications les plus populaires de l'apprentissage automatique. Ils sont utilisés pour prédire la "note" ou la "préférence" qu'un utilisateur donnerait à un élément. Presque toutes les grandes entreprises de technologie les ont appliquées sous une forme ou une autre : Amazon l'utilise pour suggérer des produits à ses clients, YouTube pour décider de la vidéo à lire lors de la lecture automatique, et Facebook l'utilise pour recommander des pages à aimer et des personnes à suivre….

8

De plus, pour certaines entreprises (Netflix et Spotify), le modèle commercial et son succès reposent sur la puissance de leurs recommandations.

Mais quels sont ces systèmes de recommandation ?

En gros, les systèmes de recommandation peuvent être classés en 3 types:

Simple recommenders : offre des recommandations générales à chaque utilisateur, en fonction de la popularité et / ou du genre du film. L'idée de base de ce système est que les films plus populaires et acclamés par la critique auront une probabilité plus grande d'être appréciés du public moyen. IMDB Top 250 est un exemple de ce système.

Content-based recommenders : suggèrent des éléments similaires basés sur un élément particulier. Ce système utilise des métadonnées d'élément, telles que le genre, le réalisateur, la description, les acteurs, etc. pour les films, afin de formuler ces recommandations. L'idée générale derrière ces systèmes de recommandation est que si une personne aime un article en particulier, elle aimera également un article qui lui ressemble.

Collaborative filtering engines : ces systèmes tentent de prédire le classement ou les préférences qu'un utilisateur attribuerait à un élément en fonction des classements antérieurs et des préférences d'autres utilisateurs. Les filtres collaboratifs ne nécessitent pas de métadonnées d'élément, contrairement à leurs équivalents basés sur le contenu.

9

Chapitre 2 : analyse et conception du projet

Dans ce chapitre on décrira l’analyse et la conception de notre projet, on commence tout d’abord par la collection des données (Scraping) en utilisant la bibliothèque BeautifulSoup du python. Ensuite on traitera la modélisation de nos données et la forme de data-set qu’on a choisi.

Comme j’ai mentionné dans le chapitre 1, il y’en a plusieurs types des systèmes de recommandation, il fallait choisit le plus convenable pour notre projet( on traitera cela en détails ).

# 2.1-Scraping :

Le terme de web scraping désigne sur un plan général une technique d’extraction automatique de contenus qui sont le plus souvent structurés, sur un ou plusieurs sites web effectuée par un tiers à l’aide programmes, bots, ou scripts spécifiques.  
Les pratiques de scraping existent depuis longtemps dans le domaine d’Internet, mais le terme a connu un regain d’usage dans le cadre du développement des techniques dites de growth hacking. Les techniques de scraping peuvent être utilisées pour alimenter des outils de veille et d’analyse (veille tarifaire e-commerce, suivi de positionnement moteur, relevé d’annonces AdWords, extraction de données sur les avis clients, etc.) ou pour constituer des fichiers dédiés à la prospection comme souvent dans le cadre du growth hacking. Des fichiers de prospection B2B « sauvage » peuvent par exemple être créés en « scrapant » LinkedIn ou le site d’un organisateur salon qui présente tous les exposants.  
Selon la forme du scraping et l’usage des données scrapées, les techniques de scraping sont souvent illégales ou tout au moins non éthiques. En effet, elles peuvent constituer une atteinte au droit de propriété intellectuelle liés aux bases de données et la collecte et l’usage des données personnelles collectées ne sont généralement pas conformes aux dispositions légales.

Pour effectuer le Scraping nous avons décidé d’écrire un code en Python en utilisant la bibliothèque Beautiful Soup.

# 2.2-DATA-SET :

Nos data-sets sont sous forme de fichiers CSV.

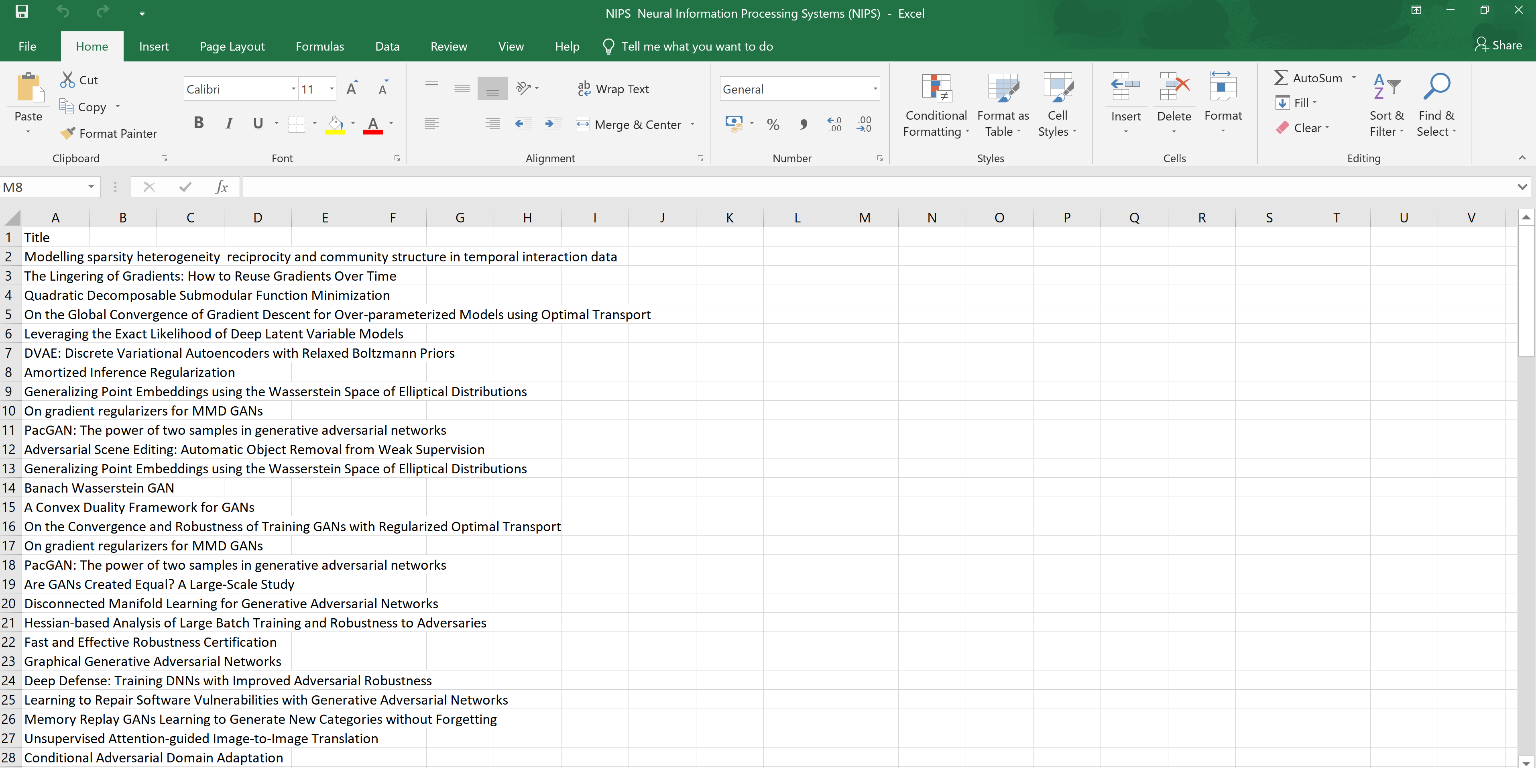
#### 2.2.1-Fichiers CSV (Comma Separated Values):

Un fichier CSV a une structure assez simple. C’est une liste de données séparées par des virgules.

Ils peuvent être plus compliqués que cela et contenir des milliers de lignes, plus d'entrées sur chaque ligne ou de longues chaînes de texte. Certains fichiers CSV peuvent même ne pas avoir les en-têtes en haut, et certains peuvent utiliser des guillemets pour entourer chaque bit de données, mais c'est le format de base.Cette simplicité est une caractéristique. Les fichiers CSV sont conçus pour être un moyen d’exporter facilement des données et de les importer dans d’autres programmes. Les données résultantes sont lisibles par l'homme et peuvent être facilement visualisées avec un éditeur de texte tel que Notepad ou un tableur tel que Microsoft Excel.

10

Ex d’un fichier CSV :



#### 2.2.2-Modélisation des données :

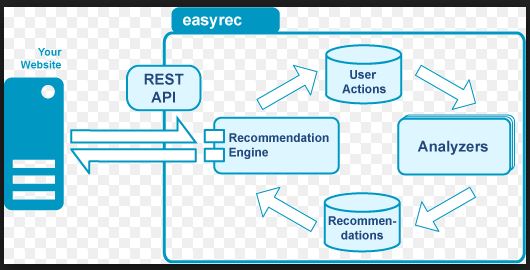
Nous avons créé une data-set centrale appelée conferences.csv : cette data-set contient les noms des conférences majeurs dans le domaine Informatique. Pour chaque conférence on associe une data-set qui contient les titres des articles acceptés dans l’édition précédente de cette conférence.

A la base de toutes ces données notre programme sera assez intelligent pouvoir de décider qu’elles sont les conférences où notre chercheur a les meilleures chances que sont article soit accepté.

Mais il faut d’abord déterminer quel type de systèmes de recommandation faut-il utiliser ?

# 2.3- Content-based recommenders :

Un content based recommender fonctionne avec les données fournies par l'utilisateur, de manière explicite (évaluation) ou implicite (en cliquant sur un lien). Sur la base de ces données, un profil d'utilisateur est généré, qui est ensuite utilisé pour faire des suggestions à l'utilisateur. Au fur et à mesure que l'utilisateur fournit plus d'entrées ou prend des mesures concernant les recommandations, le système devient de plus en plus précis.



11

Le système effectue des calculs sur le degré de similitude entre les films que vous regarder ou les articles que vous lisez et il vous recommande un article/un film similaire.

# 2.4-Pourquoi le content-based recommender system ?

Dans le cadre de notre projet, on a une data-set qui regroupe les articles des conférences, l’utilisateur donne son article et puis on calcule la similitude de cet article avec chaque article d’une conférence donnée, et donc on obtient une moyenne de similitude pour cette conférence, cette moyenne est calculée :

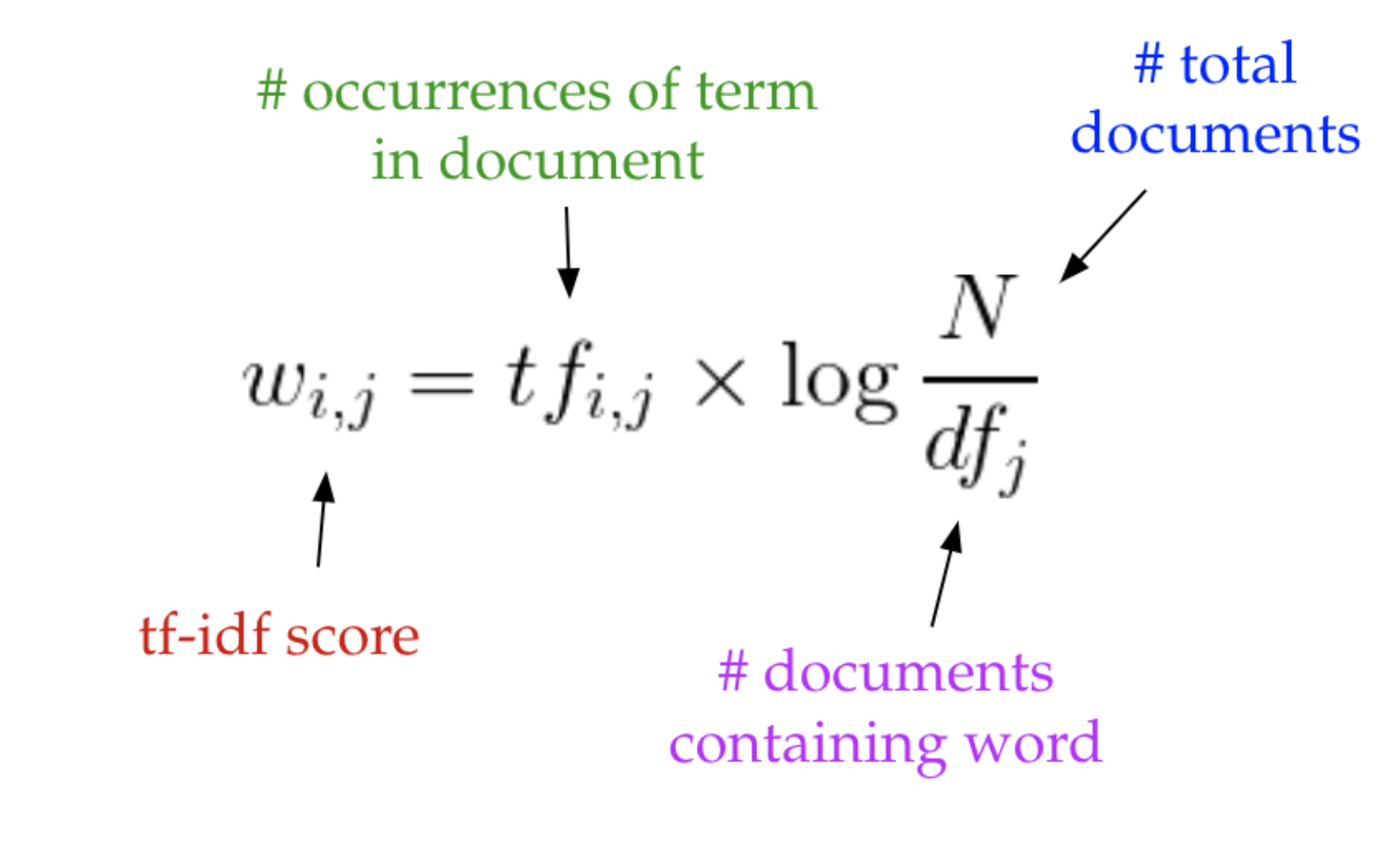
∑ (degrés de similitude de l’article de l’utilisateur avec les articles de la conférence) / (nombre d’articles)

De meme pour les autres conférences. Finalement on trie les meilleures conférences selon la moyenne de similitude.

# 2.5-LES CONCEPTS UTILISES DANS LES SYSTEMES DE RECOMMANDATION :

#### 2.5.1-TFI-IDF:

Tf-idf, or term frequency-inverse document frequency, attribue une pondération au terme j dans le document i comme suit :

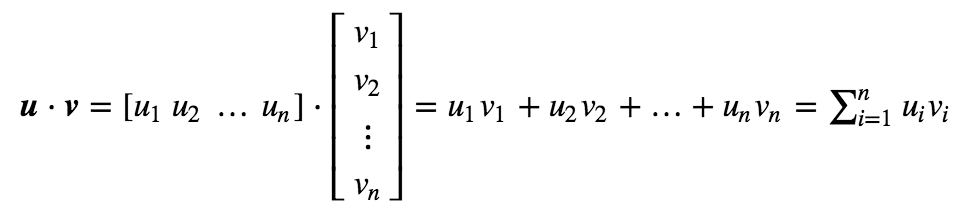


* Intuitivement, un terme a un poids important (un score tf-idf élevé) lorsqu'il apparaît fréquemment dans le *document,* mais rarement dans le *corpus*.
* Un mot peut apparaitre souvent dans un document, mais comme il est assez commun dans le reste de corpus (les autres documents), il n’aura pas un score tf-idf élevé.

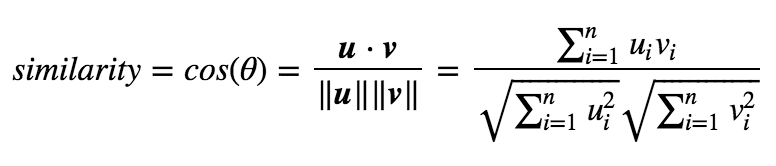
#### 2.5.2-Cosine similarity :

Nous connaissons tous les vecteurs : ils peuvent être 2D, 3D ou peu importe. Réfléchissons un instant en 2D, car il est plus facile de l’imaginer dans notre esprit, et commençons par rafraîchir le concept de produit scalaire. Le produit scalaire entre deux vecteurs est égal à la projection de l'un sur l'autre. Par conséquent, le produit scalaire entre deux vecteurs identiques (c'est-à-dire avec des composants identiques) est égal à leur module carré, tandis que si les deux sont perpendiculaires (c'est-à-dire qu'ils ne partagent aucune direction), le produit scalaire est égal à zéro. Généralement, pour les vecteurs de dimension n, le produit scalaire peut être calculé comme indiqué ci-dessous.

12



Le produit scalaire est important lors de la définition de la similarité, car il y est directement connecté. La définition de la similarité entre deux vecteurs u et v est en fait le rapport entre leur produit scalaire et le produit de leurs grandeurs.



En appliquant la définition de la similarité, celle-ci sera en fait égale à 1 si les deux vecteurs sont identiques et à 0 si les deux sont orthogonaux. En d'autres termes, la similarité est un nombre limité entre 0 et 1 qui nous dit combien les deux vecteurs sont similaires.

13

CHapitre 3 : Realisation du projet

Après avoir voir les différents concepts d’une manière théorique maintenant c’est le moment pour détailler les outils qu’ils nous ont permet de créer notre système de recommandation.

# 3.1 : le langage de programmation :



Nous avons choisi Python pour ce projet, pourquoi Python ?

Python est un langage de programmation qui existe depuis longtemps et qui est devenu très populaire. Cependant, il semblait que ce n’était pas autant sous les projecteurs que certaines des autres langues, bien qu’il soit utilisé par Google. En réalité, les développeurs Python n’étaient pas un produit très prisé, car le langage lui-même n’était pas largement utilisé comme certains autres, tels que JavaScript, Java, C ou C ++.

Et puis, ces dernières années, Python est devenu extrêmement populaire et il ne semble pas que cela ait été un hasard. En fait, Python est le langage de programmation dont la croissance est la plus rapide actuellement, et le nombre d'utilisateurs qui suivent des cours Python dans Udemy a augmenté de 34%. Mais pourquoi est-ce ? La raison est assez simple. Si nous devions analyser les sujets d’actualité de la technologie de l’informatique au cours de la dernière décennie, nous pourrions facilement comprendre que la science des données, l’apprentissage automatique et l’IA font désormais fureur. Python est idéal pour tout ce qui précède.

En plus les projets d’IA et d’apprentissage automatique peuvent être très différents les uns des autres et il est assez difficile pour un développeur d’être un homme à tout faire. C’est là que les bibliothèques entrent en jeu, car elles facilitent le codage des éléments de base et complexes. Heureusement pour les développeurs de machine learning, Python a des bibliothèques à la pelle. Numpy et Scipy sont les plus populaires, et on va citer par la suite quelques bibliothèques qui nous ont facilité la vie dans ce PFA :

14

15

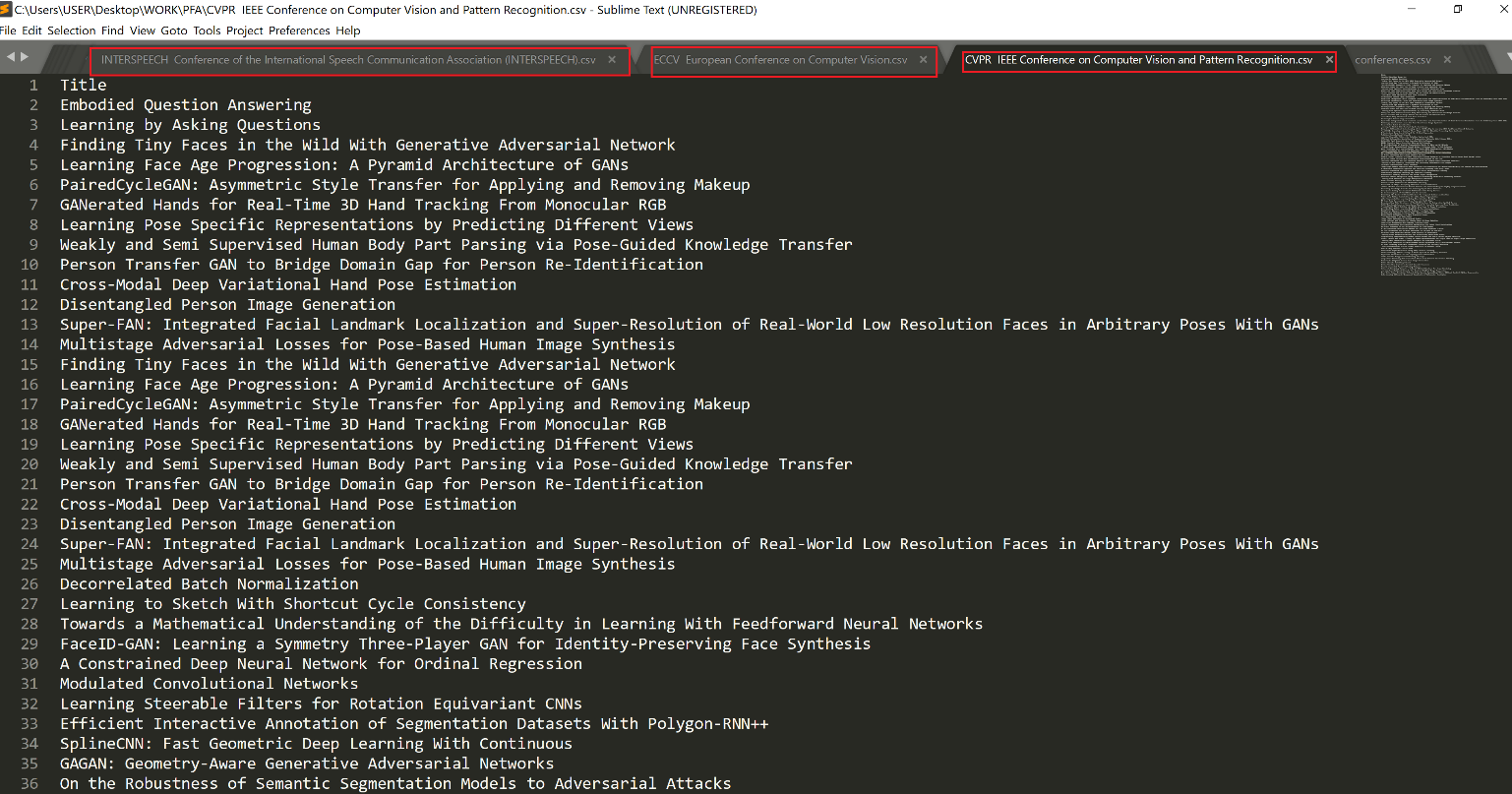
## 3.1.1-Numpy :

NumPy (Numerical Python) est une bibliothèque d'algèbre linéaire en Python. C'est une bibliothèque très importante sur laquelle presque toutes les bases de données et d'apprentissage automatique, tels que SciPy (Scientific Python), Matplotlib (bibliothèque de traçage), Scikit-learn.

## 3.1.2-BeautifulSoup :

BeautifulSoup est une bibliothèque Python permettant d'extraire des données de fichiers HTML et XML. Il travaille avec votre analyseur préféré pour vous fournir des moyens idiomatiques de naviguer, de rechercher et de modifier l’arbre d’analyse. Cela permet généralement aux programmeurs d'économiser des heures ou des jours de travail.

On l’a utilisé pour collecter les articles des conférences qui sont publiés sur différents sites, et les mettre dans des fichiers CSV sous la forme suivante :



## 3.1.3-Pandas :

Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

On a utilisé pandas pour extraire et manipuler nos données enregistrées sous forme des fichiers CSV.

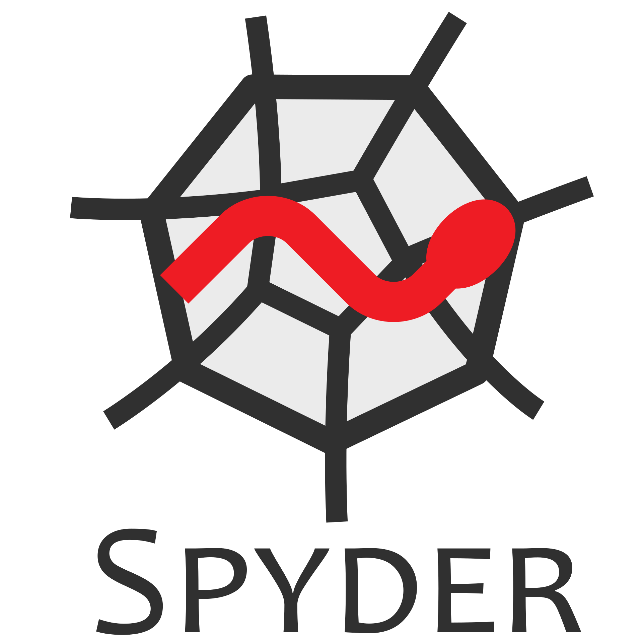
## 3.1.4-Scikit-Learn:

Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique.

15

# 3.2-Environements du développement :

## 3.2.1-Spyder :



Spyder est un environnement de développement pour Python. Libre et multiplateforme, il intègre de nombreuses bibliothèques d'usage scientifique : Matplotlib, NumPy, SciPy et IPython.

## 3.2.2-Sublime Text :



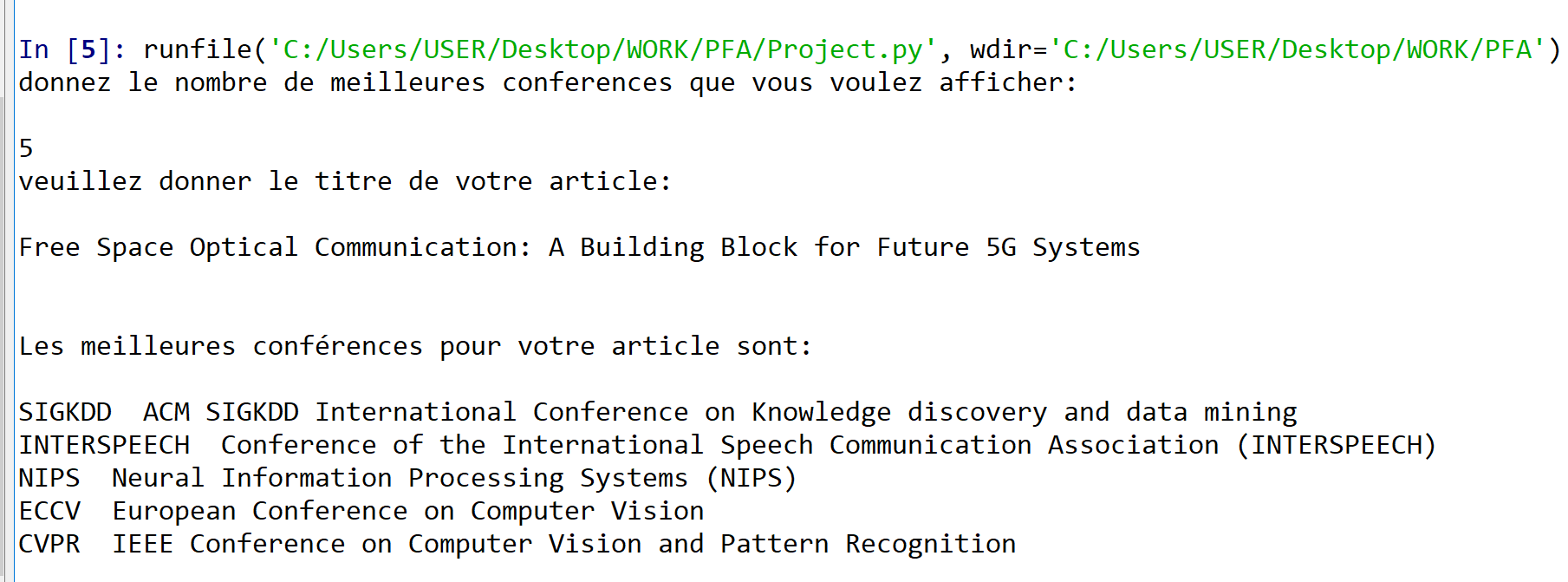
Sublime Text est un éditeur de texte générique codé en C++ et Python, disponible sur Windows, Mac et Linux.

On a utilisé sublime pour créer nos data-sets sous formes des fichiers CSV.

16

# 3.3-Realisation :

On demande à l’utilisateur de donner le nombre de conférences qu’il veut afficher, et le titre de son article. Par conséquence notre programme recommande les meilleures conférences où le chercheur a beaucoup plus de chance que son article soit accepté.



17

conclusion :

Malgré les difficultés qu’on a rencontré tout au long de la réalisation de ce projet, c’est une grande opportunité pour nous de découvrir et travailler sur des nouvelles notions telles que : le Machine Learning, Data Mining et les systèmes de recommandation qui est le sujet de notre projet.

Surtout que les systèmes de recommandation ont déclenché une révolution informatique, et sont maintenant très utilisés dans les plus grandes entreprises du monde comme : Facebook, YouTube, Netflix, Google, Amazon…….

Ce projet nous a permet d’améliorer nos compétences en programmation en Python, et de découvrir le vrai pouvoir du Python dans le domaine du ML et IA.

En substance, notre projet consiste à créer un système de recommandation (Advisor) pour les équipes de recherche, ce système a comme but : recommander les meilleures conférences où le chercheur a beaucoup plus de chance que son article soit accepté, ce qui va économiser beaucoup du temps et d’effort pour les équipes de recherche.

18

Bibliographie

<http://www.guide2research.com/topconf/machine-learning>

<https://www.datacamp.com/community/tutorials/recommender-systems-python?utm_source=adwords_ppc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=1t1&utm_creative=332602034358&utm_targetid=aud-299261629574%3Adsa-473406569915&utm_loc_interest_ms=&utm_loc_physical_ms=1009984&gclid=CjwKCAjw-ZvlBRBbEiwANw9UWlAmpQMqL14LFgiS1kfbanUkpnMjfdrZbBILUUvrLJLxvFTDBxVF1hoCrlUQAvD_BwE&fbclid=IwAR3bh9EDlk5YAIeljUe9s5oF0Rn6_PRrmiWY7-fnclzFSkb7BGNm1EA4cbw>

<https://medium.com/dunder-data/minimally-sufficient-pandas-a8e67f2a2428>

<https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05>

<https://medium.freecodecamp.org/how-to-scrape-websites-with-python-and-beautifulsoup-5946935d93fe?fbclid=IwAR02fh2UHj71MP7Qq28OSKFIOodb1Ym6FdktnYHBcfhUY095O8t_LMTD38c>

<https://stackoverflow.com/>

<https://towardsdatascience.com/lets-talk-about-numpy-for-datascience-beginners-b8088722309f>

19