NOTE DE CADRAGE

Définition des enjeux :

Visualiser rapidement le retour de l’expérience client sur un produit afin d’être plus compétitif.

Définition des objectifs :

L’objectif du programme est d’analyser les commentaires du site marchand Zalando (voir visuel page suivante) en fournissant les résultats d’une analyse statistique et d’une analyse « sémantique ». Il s’agit de révéler les mots qui ressortent le plus et de générer la proportion de contenus positifs et négatifs. Le public cible est constitué des marques de prêt à porter et de chaussures dont les produits sont en vente sur le site et qui auraient envie de voir rapidement les avis concernant leurs produits, suite à une analyse du contenu des commentaires.

L’étude des sentiments est basée sur une analyse « sémantique », à partir d’un dictionnaire. A la différence d’une analyse quantitative dont les notes peuvent être faussées, les avis peuvent mieux être étudiés et pris en compte pour connaître les éventuels points forts et points faibles du produit en question dans une optique d’amélioration.

Définition des priorités :

Priorité 1 : montrer visuellement les mots qui ressortent le plus dans une page web.

Priorité 2 : analyser les sentiments (positifs ou négatifs) des avis des consommateurs en se basant sur la sémantique.

Priorité 3 (optionnel): gestion d’une mauvaise entrée d’url (autre que Zalando).

Définition de l’organisation :

Les membres de notre groupe constitué d’Anne Bauer, de Madeleine Bimai et de Virginie Rieber, s’occuperont de toutes les phases du projet Python, avec une répartition des tâches dans chaque phase. Les tâches simples et moyennes seront réparties entre nous. Enfin nous effectuerons les tâches complexes en commun.

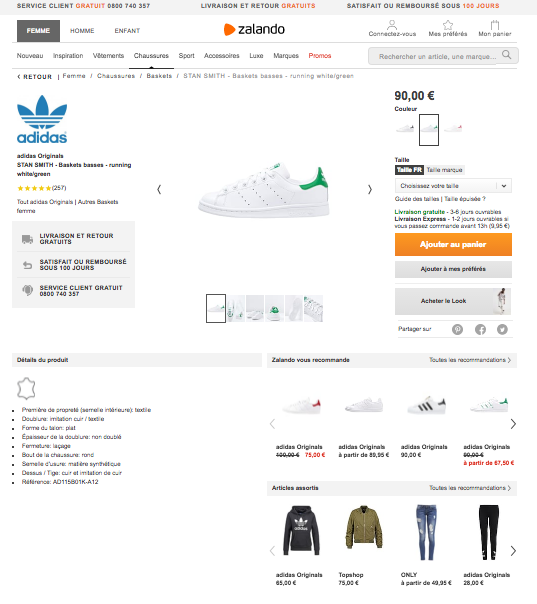
Définition des principaux risques :

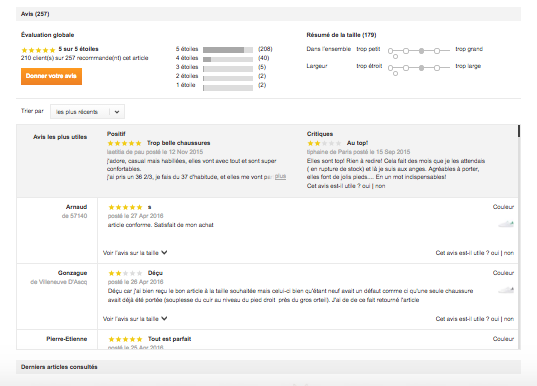
Une analyse importante est à effectuer en amont pour trouver les ressources : librairies (NumPy, NLTK). Il s’agit de trouver des ressources disponibles afin de maximiser l’efficacité du projet et de ne pas réinventer ce qui existe déjà.

Le dictionnaire utilisé dans l’analyse risque de ne pas être exhaustif.

Les risques sont liés au langage naturel qui comporte des ambiguïtés, l’ironie ou encore les fautes d’orthographe ou le langage sms. Sont à prendre en compte également l’impact de la perte de sens dans l’analyse due au contexte (tokenisation : traitement des mots de manière indépendante). Il faudra aussi porter une grande attention à la fiabilité de l’analyse due à la proportion de données que l’on conserve et qu’on l’on jette en fonction du dictionnaire (sujets, articles, etc.).

D’autres risques sont liés au temps, à la charge de travail du master, à l’alternance des périodes à l’université et en entreprise, et donc à la possibilité que les membres du groupe puissent se rencontrer.





Bibliographie :

[1] S. Bird, E. Klein, et E. Loper, *Natural language processing with Python: [analyzing text with the natural language toolkit]*, 1. ed. Sebastopol, Calif.: O’Reilly, 2009.

[2] « Natural Language Toolkit — NLTK 3.0 documentation ». [En ligne]. Disponible sur: http://www.nltk.org/. [Consulté le: 29-avr-2016].

[3] « NumPy — Numpy ». [En ligne]. Disponible sur: http://www.numpy.org/. [Consulté le: 29-avr-2016].

Annexes :

Processus (et estimation de la complexité des tâches)

s :simple

m : moyen

c : complexe

1. Analyse amont/Création des ressources :
   * Recherche documentaire (librairies python) : c
   * BDD : m
   * 3 dictionnaires : m
2. Création interface : m
3. Connexion / Chargement des données

* Lecture de l’URL : s
* Connexion à l’URL : s
* Identification de la zone de commentaires : s
* Extraction du contenu : s
* Comparaison des données à un dictionnaire (tableau ou autre) et tri (élimination des verbes, sujets, etc.) : c
* Stockage du contenu dans un tableau 1 : s

1. Analyse quantitative

* Comptage du nb d’occurrence de chaque élément distinct du tableau 1 et création d’un tableau 2 avec le nb d’occurrence et les mots associés : m
* Tri du tableau 2 dans l’ordre décroissant : s

1. Production graphique

* Chargement du tableau 2 : s
* Définition des repères du graphique et remplissage du graphique avec les données du tableau 2 : c

1. Analyse sémantique

* Copie tableau 1 dans tableau 1 bis : s
* Comptage du nb d’éléments du tableau 1 bis: s
* Comparaison de chaque élément du tableau 1 bis avec le dictionnaire « + »: m
* Copie des éléments dans un tableau 3 si présence dans le premier dictionnaire et suppression de l’élément dans le tableau 1 bis : s
* Comparaison des éléments restants du tableau 1 bis avec un deuxième dictionnaire : m
* Copie des éléments dans un tableau 4 si présence dans le dictionnaire « - » et suppression de l’élément dans le tableau 1 bis : s
* Comptage du nb d’éléments dans le tableau 1 bis, 3 et 4 : s
* Calcul du pourcentage : s

1. Production graphique « sentiment »

* En fonction du pourcentage, génération d’un diagramme de type jauge (histogramme horizontal) : c