Extraction d’opinion sur des avis utilisateurs de SensCritique

Travail sur les séries

Groupe : Erna, Sohaila, Nidia, Justine, Mathilde

Introduction

Ce travail expose la démarche suivie pour une analyse automatique de textes d'opinion. Il s'agit surtout d'une étude de type descriptive où l'on compare les distributions de certaines unités textuelles au sein des différentes parties du corpus, mais quelques propositions interprétatives sont avancées pour certains points.

La section suivante explique les objectifs visés par ce travail. La section "Corpus" décrit les données sélectionnées pour l'analyse et les prétraitements nécessaires pour sa manipulation automatique. Les sections suivantes détaillent les différentes expériences effectuées sur le corpus à l'aide de TXM et la prise en compte des caractéristiques extraites pour l'entraînement d'un classifieur. De brèves analyses de nos résultats sur mènent finalement à nos conclusions.

Objectifs

Notre objectif était d'analyser lexicométriquement un corpus et les éléments linguistiques déterminants en fonction de l'avis exprimé. Cela relève donc de l'analyse de sentiment mais nous présentons ici une analyse exploratoire des éléments notables.

Nous avons donc mis en lumière pour chaque partition du corpus ses caractéristiques linguistiques propres et les avons extraites afin de les utiliser pour mettre en place un système permettant de détecter automatiquement la teneur de l'opinion exprimée.

Corpus

Extraction du corpus

Le corpus est extrait du site SensCritique qui regroupe des critiques d'oeuvres diverses, autant des films que des albums musicaux en passant par les bandes dessinées. Chaque commentaire est une critique sur une oeuvre en particulier et est accompagné d'une note attribuée par l'auteur de la critique. Dans l'objectif initial de faire de l'apprentissage automatique à partir de ce corpus, l'utilisation de ce site permettait d'avoir un commentaire et en même temps la note pour vérification des résultats obtenus.

Le script d'extraction est disponible dans le git du projet.

Description

Le corpus a été extrait au format XML et JSON et contient les informations suivantes :

- le titre de l'oeuvre
- la critique
- le pseudo de l'auteur

- la date
- la note attribuée.

Extrait du fichier sur les séries (les critiques ont été nettement raccourcies par soucis de lisibilité, nous souhaitons ici montrer la structure des données et les informations fournies) :

```
<?xml version='1.0' encoding='UTF-8'?>
<corpus>
  <oeuvre titre="Powers">
    <critique auteur="Marvelll" date="2015-03-13 22:04:16+01:00" note="7">
Critique des trois premiers épisodes :
Powers est une série bizarre. C'est moche, mal réalisé et le support technique est catastrophique (coupures réseaux, application gratuite planquée), mais bizarrement, je l'aime bien. Il faut dire que les acteurs ne sont pas mauvais et l'intrigue donne globalement envie d'en découvrir plus, sans compter sur un univers solide, même si peu original. [...]
    </critique>
    <critique auteur="BlackLemmy" date="2016-09-02 04:03:09+02:00" note="6">
Powers fait partie de ces séries dont tu te demandes un peu ce qu'il s'est passé pour qu'elles voient le jour telles quelles.
 [...] Encore un coup d'épée dans l'eau pour le genre Fantastique du "Mais comment vivre normalement quand on est un monstre ?" (voir la liste suivante :
http://www.senscritique.com/liste/Qui_a_dit_que_les_series_Fantastique_c_etait_toujours_la_mem/141225#page-1/ )
</critique>
    <critique auteur="Walhan" date="2015-08-07 10:03:03+02:00" note="6">
Une nouvelle série de super héros (adaptation d'un comics), dont les personnages principaux n'ont pas vraiment de pouvoirs...
 Un pari osé mais qui a déjà fonctionné.[...]
Retrouvez la fiche sur les Découvertes de Walhan
</critique>
```

Pré-traitements et partition

Nous avons travaillé sur les séries en particulier et avons donc récupéré les critiques de séries uniquement. Puis le corpus a été partitionné selon les notes dans 3 catégories :

- négatif : notes de 1 à 3
- neutre : notes de 4 à 7
- positif : notes de 8 à 10

Méthode

Analyse Exploratoire

Méthode

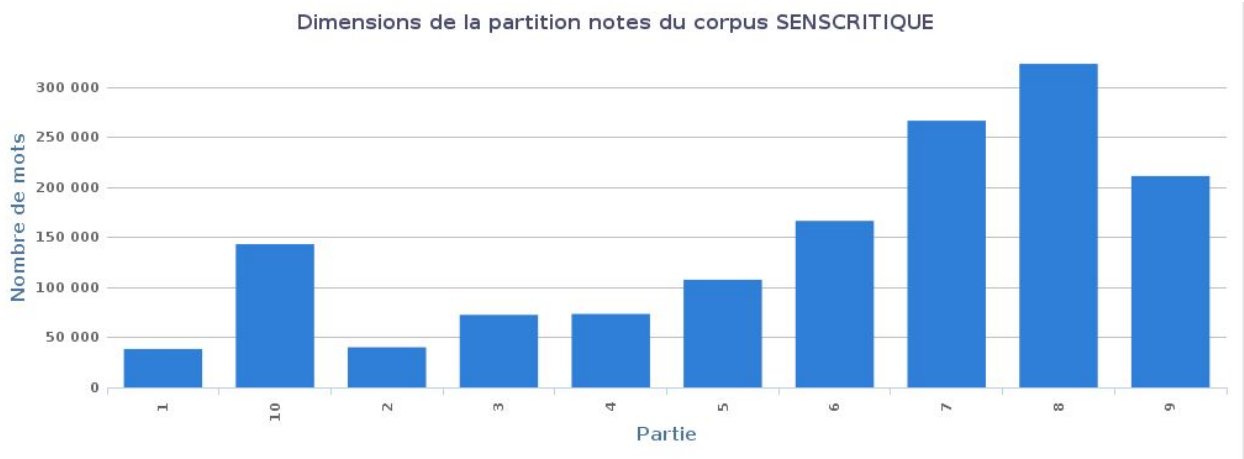
Afin de repérer les éléments discriminants pour chaque sous-partition, nous avons réalisé une analyse exploratoire. Cette analyse a été réalisée avec le logiciel de textométrie

TXM. Nous avons analysé les différentes vues du corpus qu'offre TXM et effectué des recherches sur les occurrences des mots et leurs co-occurrents ainsi que plusieurs calculs de spécificités.

Remarques générales

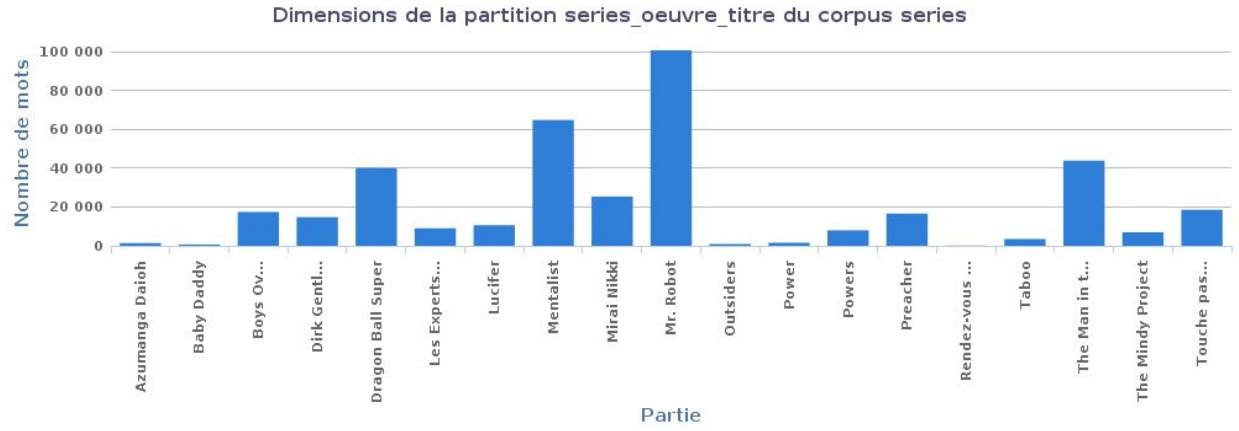
Taille et répartition du corpus

Avant toute chose, nous remarquons que le nombre de mots est assez inégal entre les critiques positives, neutres et négatives.



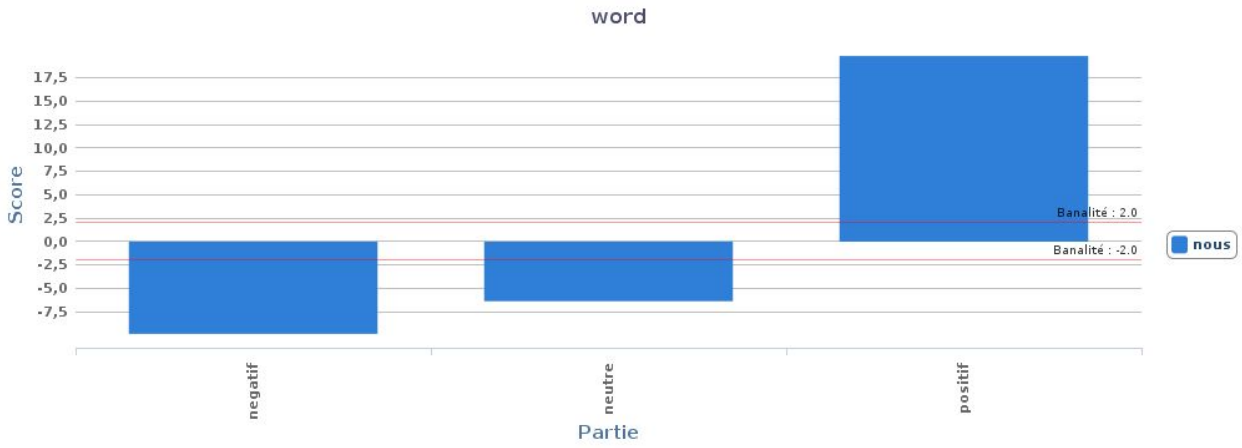
On peut penser que cela montre que les critiques positives sont plus verbeuses que les critiques négatives.

On note aussi que toutes les séries n'ont pas le même nombre de critiques :



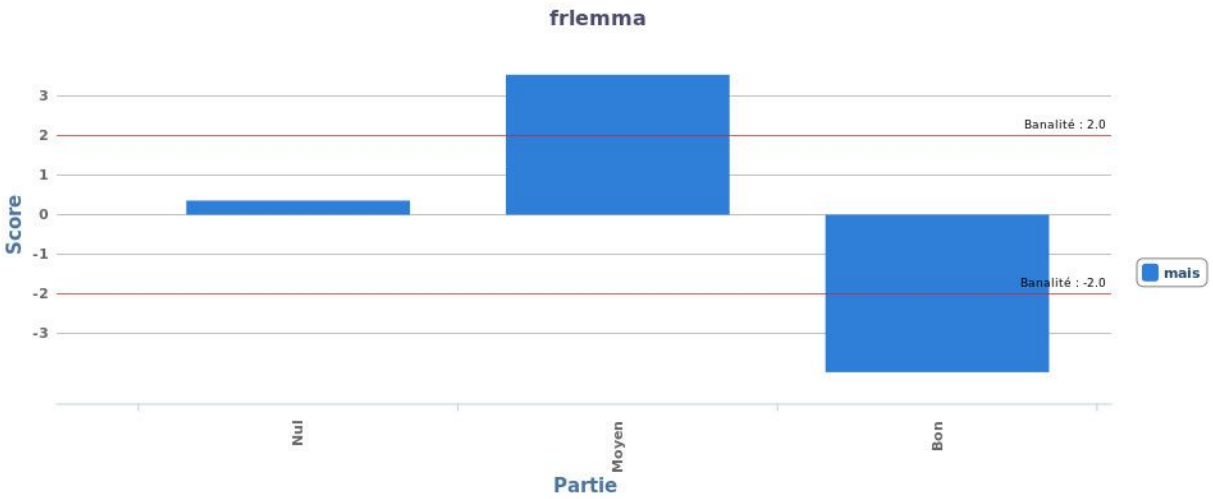
Répartition générale des mots

Afin de mettre en avant les éléments discriminants, nous avons étudié les mots les plus spécifiques de chaque partition. On constate que "nous" est fortement représentatif du corpus positif.



Les connecteurs et marques d'opposition ou de concessions sont principalement utilisé dans les critiques neutres ou positives, ici deux expressions et les notes des critiques contenant le plus d'occurrences :

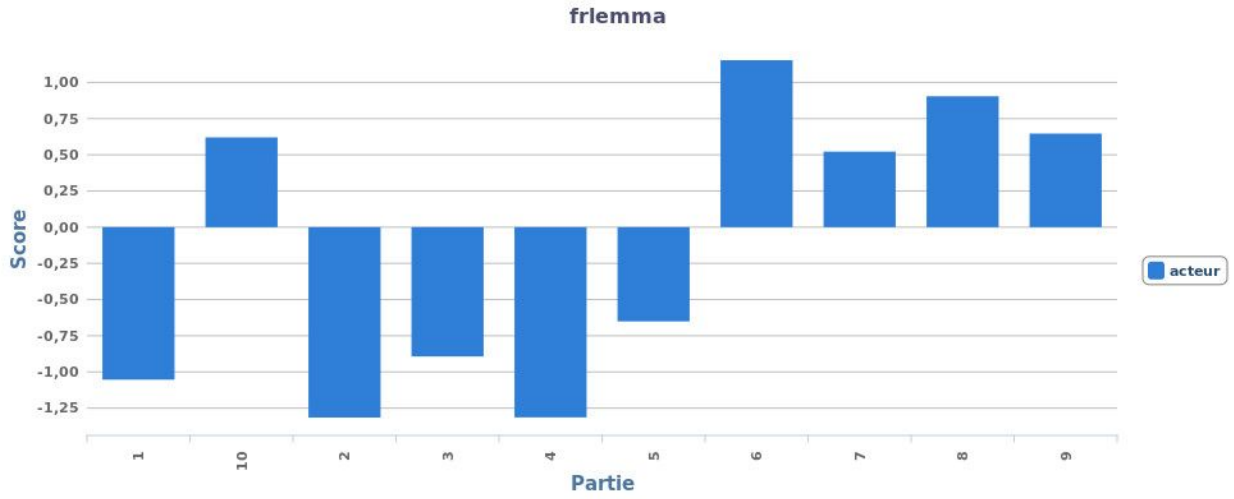
- mais : 4, 5, 6, 7
- alors que : 7,8,9,10

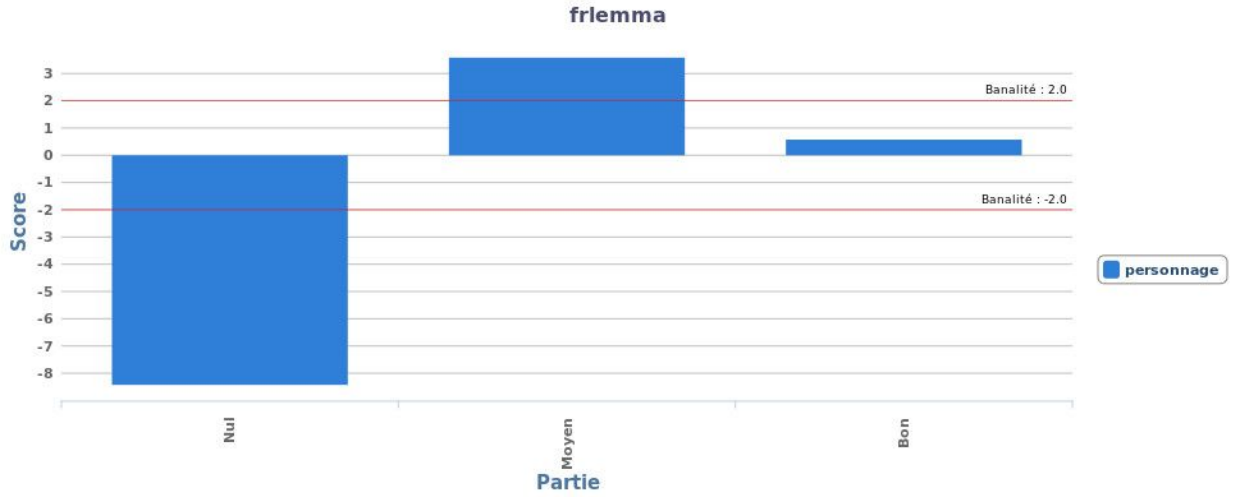


En plus d'avoir noté des critiques plus verbeuses, cela confirme l'idée d'un discours plus construit dans les avis positifs.

Les personnages et acteurs

L'analyse des mots associés aux personnages des séries montre que les utilisateurs ont tendance à parler des "personnages" si la série obtient des notes entre 4 et 7 et des "acteurs" si la série est considérée comme bonne (notes entre 8 et 10).





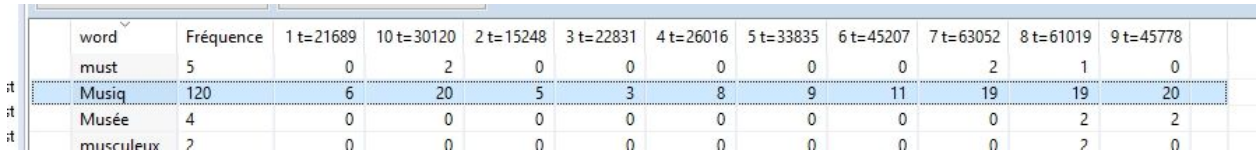
On constate avec la recherche d'expressions régulières que la recherche de type EN+est+ADJ (entité nommé suivie de "est" suivi d'un adjectif) renvoie des phrases sur les acteurs et personnages et on obtient principalement des adjectifs positifs. C'est une spécificité des critiques positives se référant aux acteurs.

La liste des cooccurrents de "principal" confirment qu'il se rapporte presque exclusivement au "personnage" ou à l'"acteur" principal et les premiers cooccurrents sont des mots positifs. Cela est a mettre en relation avec le fait que notre corpus contient plus de critiques positives et qu'elles sont plus verbeuses, ce n'est donc potentiellement pas un élément fortement discriminant.

cooccurrences de 'principal' : personnage, acteur, jeu, grand, bonne, super, rôle, intérêt, charisme.

Musique et Bande son

Le sujet de la musique revient aussi principalement dans les critiques avec les meilleures notes, dans notre cas à partir de la note 6, et le maximum d'occurrence se trouve dans les critiques ayant donné la note 10.



word	Fréquence	1 t=21689	10 t=30120	2 t=15248	3 t=22831	4 t=26016	5 t=33835	6 t=45207	7 t=63052	8 t=61019	9 t=45778
must	5	0	2	0	0	0	0	0	2	1	0
Musiq	120	6	20	5	3	8	9	11	19	19	20
Musée	4	0	0	0	0	0	0	0	0	2	2
musculeux	2	0	0	0	0	0	0	0	0	2	0

Parmi les cooccurrents on retrouve "parfaitement ,"belle", "excellente". La musique est liée à l'ambiance et au suspense la plupart du temps, la musique du générique revient aussi et semble être un critère dans les critiques.

Adaptation

La disponibilité d'un nombre considérable de segments analysables est un prérequis fondamental pour les analyses textuelles automatiques. Ainsi, le choix initial du descripteur - décor/costumes - a été modifié en raison des faibles occurrences dans le corpus:

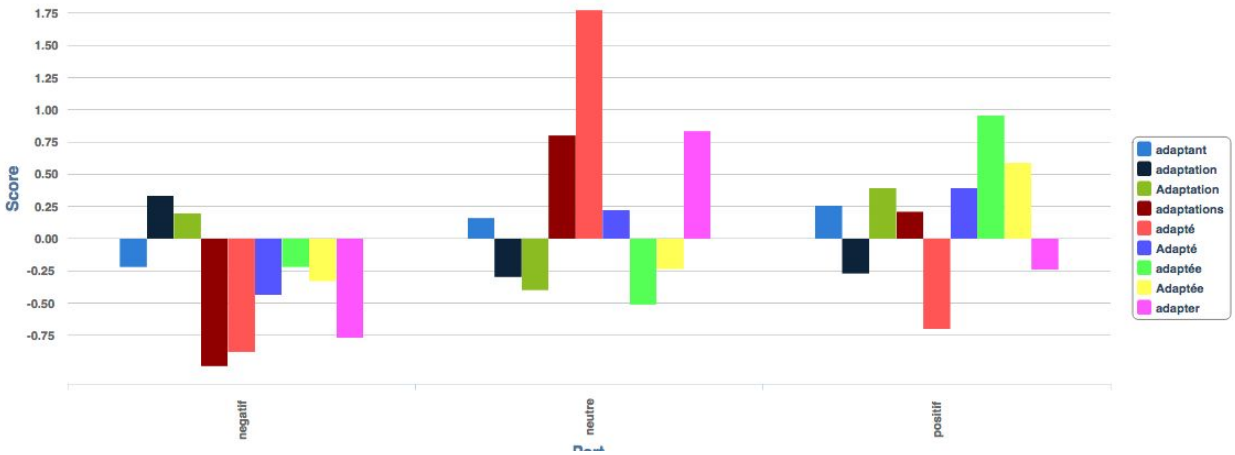
```
Concordance of <"costume"> in corpus CORPUS
Found 5 occurrences
Concordance of <"décor"> in corpus CORPUS
Found 11 occurrences
```

Une recherche rapide du descripteur "adaptation" et des mots sémantiquement liés (oeuvre, livre, manga, etc) a donné un nombre d'occurrences plus élevé. Il a donc été décidé de travailler sur ce sujet.

```
Concordance of <"oeuvre"> in corpus CORPUS
Found 41 occurrences
Concordance of <"adaptation"> in corpus CORPUS
Found 55 occurrences
Concordance of <"adapter"> in corpus CORPUS
Found 7 occurrences
Concordance of <"roman"> in corpus CORPUS
Found 36 occurrences
```

```
Concordance of <"livre"> in corpus CORPUS
Found 62 occurrences
Concordance of <"BD"> in corpus CORPUS
Found 15 occurrences
Concordance of <"bd"> in corpus CORPUS
Found 10 occurrences
Concordance of <"manga"> in corpus CORPUS
Found 103 occurrences
```

Le calcul de spécificité pour les lemmes "adaptation"/"adapter" évidence une certaine préférence pour les commentaires positifs par rapport aux négatifs, mais globalement, ces lemmes sont propres des commentaires neutres :



La recherche d'expressions qu'on attendait trouver à propos des adaptations n'a pas donné de résultats satisfaisants :

```
Concordance of <[word="inspiré"][word="PRP"] [frpos="NAM"]> in corpus CORPUS
Done: no result
```

Cela peut effectivement signifier que ces expressions ne se trouvent pas dans le corpus mais il peut être la conséquence d'un mauvais étiquetage de Treetagger. Une recherche plus générale par expression régulière `[word="inspiré"][frpos="PRP"][word=".*"]` rend les concordances suivantes :

```
sein de cette mixture un groupe de  inspiré de   " Community ". Mais la recette ne fait le chef personnages                                    cuisinier

la salle de bain en français, est   inspiré d'   un épisode de Doctor Who écrit par Douglas Adams himself, la

chaque coup. Sam Esmail dit avoir   inspiré par  le film « Taxi Driver » mais ce qui saute le plus été

...) Un polar efficace et piquant,  inspiré par  une bande dessinée. A voir en VO bien sûr....
```

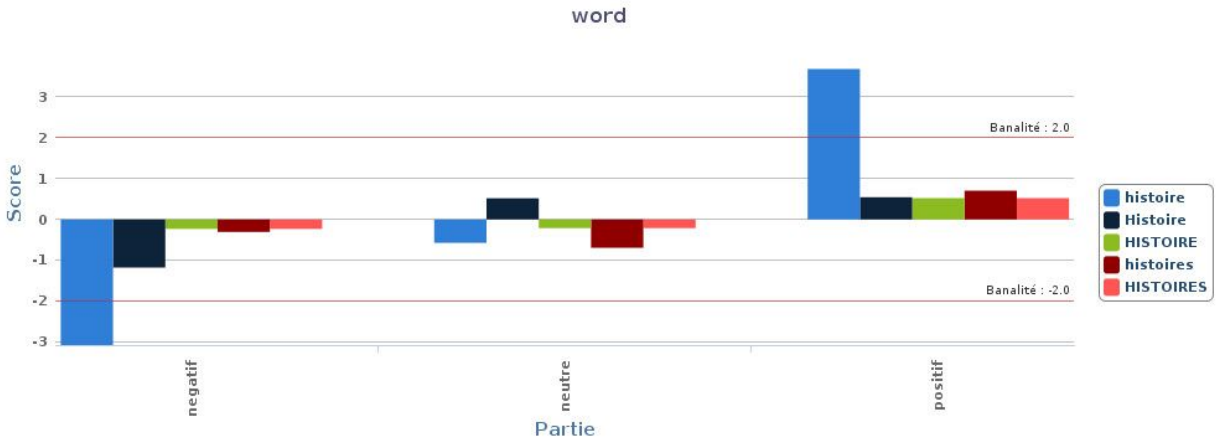
Effectivement, la première concordance doit être mal étiquetée. Dans les trois autres cas, nos recherches précédentes étaient trop restrictives, les internautes préfèrent préciser plus les références ("le film Taxi Driver" au lieu de tout simplement Taxi Driver). Cela correspond à une tendance générale des internautes à hyperpréciser le sujet dont ils parlent, probablement car ils cherchent à se situer comme des experts qui connaissent les sources d'inspiration des séries dont il est question :

```
(je précise que j'ai vu la version non-censurée ainsi que les OVAs)
… est inspiré d'un épisode de Doctor Who écrit par Douglas Adams himself
Ce qui m'intéresse ici c'est la version coréenne diffusée en 2009. La série compte 25 épisodes d'une heure…
```

L'expression `[word="version"][frlemma="de"][frpos="DET"]?[frpos="NOM"]` n'a pas donné de résultats. Par contre, `[word="version"][frpos="ADJ"]` a rendu 33 concordances dans la plupart des cas pertinentes ("version japonaise de ce manga…", "Cette version coréenne n'arrive pas à la cheville", etc).

Taglines (phrases d'accroches)

Histoire et scénario



Les différentes occurrences du mot “histoire” sous plusieurs formes montrent aussi une plus grande utilisation dans les critiques positives. Les cooccurrents mettent en avant une histoire “vraie” ou l’histoire “principale” et le “déroulement” de celle-ci, vient ensuite l’histoire “intéressante”, “prenante” ou bien “ficelée”. Enfin on retrouve l’idée d’une histoire “originale”. À peine plus loin on retrouve les mots “confuse”, banale”, “simple” mais dans le cas de “simple”, l’expression est presque systématiquement suivie d’une concession :

“Une histoire simple mais belle”

“Une histoire simple mais efficace”

“Une histoire simple sans pour autant être simpliste”

De même “l’histoire [...] banale” est pourtant “si réelle”, “rendue spectaculaire par les acteurs” et “bien structurée”.

Du côté des critiques négatives, l’histoire est “inintéressante”, “ennuyeuse” et fait trop dans la “simplification”.

Critiques négatives

Dans les critiques négatives, à l’opposé du “nous” des critiques positives, ce sont les pronoms de la première personne qui ressortent, ainsi on retrouve “me”, “je”, “j’”.

On remarque aussi le pronom personnel “se” très utilisée pour “se retrouver”, “se permettre”, “se battre”, “se moquer”, “se compter”, “se laisser” …

Entraînement d'un classifieur

À partir des traits discriminants que nous avons pu mettre en lumière lors de l'analyse exploratoire, nous avons mis en place une extraction automatique de plusieurs éléments pour entraîner un système d'apprentissage automatique.

Nous avions effectué deux extractions à partir de SensCritique, ce qui nous a permis de constituer un corpus de test et un corpus d'entraînement. Ces deux fichiers se trouvent dans le dossier "Corpus" du projet sur GitHub.

Choix des traits

Pour la détection des critiques positives, nous avons décidé d'utiliser le terme "nous" qui est fortement discriminant. Pour les critiques un peu plus neutres, c'est l'utilisation de connecteurs comme "mais" et "alors que".

Nous avons aussi constaté la différence de longueur entre les critiques positives et négatives, nous avons donc ajouté comme trait le nombre de mots.

Enfin, le corpus montre une certaine différence au niveau de l'utilisation de la ponctuation, dans l'idée que les critiques positives sont plus verbeuses et plus construites, l'utilisation de la virgule est plus fréquente, à l'inverse, les critiques négatives étant plus spontanées, c'est le point d'exclamation qui apparaît plus souvent. Pour prendre cela en compte, on utilise le nombre de points d'exclamation et le nombre de virgules comme traits supplémentaires.

Scripts

Notre projet est constitué de plusieurs scripts pour chaque étape de la chaîne de traitement :

- **extract_critique.py** : récupération des critiques et passage du format xml au format tabulaire
- **main.py** : script principal faisant appel à tous les autres scripts : nettoyage du corpus, , extraction du texte des critiques, création du tabulaire pour le corpus d'entraînement
- **corpus_test.py** : création et mise au format tabulaire du corpus de test
- **getInfos.py** : récupération des informations sur les critiques (nombre de mots, nombre de points d'exclamation ...)
- **nettoyage.py** : scirpt pour le nettoyage du corpus (enlève les if(!), intègre des espaces avant et après les ponctuations ...)

Wapiti

Pour l'utilisation de Wapiti, deux fichiers sont nécessaires :

- **config.crf** : fichier de configuration pour wapiti, prend en compte les 9 features du modèle, décrits dans le fichier
- **wapiti.sh** : lancement de wapiti (entrainement et application du modèle)

Wapiti fournit un fichier de modèle et un fichier de sortie qui contient 10 colonnes, la dernière colonne contenant la prédiction de wapiti.

Évaluation

Pour l'évaluation, nous avons réalisé le script suivant :

- **conllEval.py** : évaluation

Ce script permet de comparer les deux dernières colonnes du ficheir de sortie de wapiti (sortie.wap) donc la réponse attendue et la réponse prédite par Wapiti, à partir de là, nous avons pu obtenir une évaluation de nos résultats :

Exactitude : 32.8358208955%			
	Positifs	Neutres	Négatifs
Rappel	**0.883177570093**	0.778761061947	0.245398773006
Précision	0.615635179153	0.709677419355	**0.833333333333**

On obtient une exactitude assez basse, nous avons utilisé peu de traits, ils sont loin d'être assez nombreux pour permettre des résultats plus satisfaisants.

Au niveau du rappel et de la précision, on constate que les résultats sont convenables mais insuffisants au niveau des critiques négatives. Le rappel montre que les critiques positives sont bien reconnues, les traits choisis sont pertinents. Nous nous sommes principalement appuyé sur les traits des critiques positives pour notre classifieur, ce qui peut expliquer des résultats passables pour les critiques négatives.

En matière de précision, les résultats sont meilleurs pour les critiques négatives principalement en raison de leur faible nombre. En effet les critiques positives sont plus nombreuses, les traits les concernant sont donc beaucoup plus performants.

Les résultats globalement obtenus montrent malgré tout que les traits utilisés sont discriminants, assez pour permettre une première classification relativement intéressante.

On peut donc dire que notre corpus était trop inégal pour permettre une extraction assez pertinente des traits pour les critiques négatives mais nos résultats sont convenables pour les critiques négatives. Les traits linguistiques déduits de

l'analyse exploratoire montrent ici leur intérêt même si les traits de surface notamment le nombre de mots reste un trait important.

Autres traits envisagés

Nous avons aussi considéré d'autres traits comme le nombre de "je" qui est significatif des critiques négatives, voici les résultats obtenus :

Exactitude : 34.4941956882%			
	Positif	Neutre	Négatif
Rappel	0.855140186916 (-)	0.761061946903(+)	0.245398773006(-)
Précision	0.598039215686 (+)	0.690763052209(-)	0.833333333333(-)

Comme on le voit, la prise en compte de la présence du "je" permet de gagner deux points en exactitude, une légère amélioration. En effet, "je" était un élément caractéristiques des critiques négatives mais il ne semble pas être suffisamment discriminant pour vraiment faire une différence.

De même, la prise en compte de la présence du mot "musique" fait passer l'exactitude à 35.8208955224%, ce qui nous permet de dire que ce n'est un thème suffisamment discriminant pour améliorer la détection automatique des opinions.

Conclusion/Perspectives

La prochaine étape serait de prendre en compte un étiquetage des critiques et une recherche de patrons morphosyntaxiques comme ceux que nous avons mis en avant dans les parties précédentes.

Dans ce projet nous avons entamé une étude sémantique d'un corpus textuel à travers une démarche mixte et complémentaire : des requêtes simples et avec des regex nous ont permis de détecter des expressions pertinentes pour entraîner un algorithme de classification automatique des critiques selon le type de valoration. Autrement dit, nous avons adopté un premier approche symbolique pour délimiter l'action d'une deuxième étape de traitement par apprentissage automatique.

Des prolongements possibles de ce travail seraient:

- Effectuer plus de requêtes pour d'autres descripteurs nous servant des étiquettes morphosyntaxiques pour recueillir plus d'expressions pertinentes pour la classification;
- Réaliser un classement automatique à l'aide d'un autre algorithme de Wapiti (par exemple sgd-l1) pour comparer les résultats;
- Confronter la performance de Wapiti avec celle de Weka;
- Analyser plus en détail les cooccurrences des termes choisis pour le classement et effectuer une lecture des commentaires de chaque partition pour vérifier leur pertinence et approfondir notre étude.

Les techniques de traitement automatique de texte testées dans ce projet ont montré une efficacité satisfaisante et représentent une grande économie de temps pour l'analyse du corpus. Cependant une analyse interprétative tenant en compte des facteurs contextuels s'impose pour vraiment approfondir dans le sens des textes.