



M1 INFORMATIQUE AIGLE

HMIN232M

MÉTHODES DE LA SCIENCE DES DONNÉES

Classification de documents d'opinions

Bachar RIMA
Tasnim SHAQURA
Emile YOUSSEF

23 avril 2019

Table des matières

1	Introduction	1
2	Pré-traitements des données	2
2.1	Préparation à la tokenisation	2
2.2	Tokenisation et normalisation	4
3	Apprentissage du modèle	7
3.1	Vectorisation	7
3.2	Feature-Engineering	7
3.3	Cross validation et calibrage des hyperparamètres	7
3.4	Sauvegarde du modèle	7
4	Optimisation	8
4.1	WordCloud	8
4.2	Traitement de l'ironie	8
5	Conclusion	9
	Annexes	10
	Annexe A Snapshots de pré-traitement	11

Chapitre 1

Introduction

Aujourd'hui, on trouve sur Internet un grand nombre de données quantitatives, mais aussi et surtout, des données textuelles qualitatives. Ces données se divisent en deux domaines principaux : faits et opinions. D'un côté, les faits reposent sur l'objectivité et la fiabilité des données et de l'autre, les opinions représentent les sentiments de leurs auteurs.

De plus en plus de sites web offrent la possibilité aux clients de laisser leurs avis sur leurs produits. Parfois l'avis d'un client est accompagné par une note facilement interprétable en tant que mesure de satisfaction du client. Cependant, ce n'est pas toujours le cas. Une fouille des avis subjectifs s'avère ainsi essentielle pour apporter des informations supplémentaires sur les sentiments des clients, et par conséquent en tirer les défauts et les avantages des produits concernés. Ceci permet aux vendeurs d'avoir une idée beaucoup plus précise des attentes de leur clientèle afin de mieux y répondre.

L'analyse du sentiment de l'avis d'un internaute se fait de manière naturelle pour un être humain. Cependant, l'explosion actuelle des données ne lui permet plus de les analyser de la même manière. Pour résoudre ce problème, le traitement automatique du langage naturel et la classification des textes sont des tâches inévitables à effectuer par n'importe quelle machine souhaitant traiter des données volumineuses de sentiments. Ces approches permettent de récolter des statistiques sur un texte et de reconnaître et filtrer des motifs y inclus afin de révéler son orientation.

L'objectif de ce projet est de créer une intelligence artificielle qui permettrait, à partir d'une phrase, de reconnaître la polarité de celle-ci : est-ce un avis positif ou bien négatif ? Comment lui apprendre à reconnaître la négation ou bien encore plus complexe, le sarcasme ? Nous essaierons d'y répondre par la suite.

Chapitre 2

Pré-traitements des données

Dans le cadre de l'analyse des sentiments, le pré-traitement des documents à classifier est fondamental afin d'obtenir un classifieur généralisable, performant et assez robuste. On distingue ainsi deux grains de pré-traitement :

document : traitement de l'ensemble des tokens d'un document collectivement.

token : traitement de chaque token d'un document individuellement.

À cet effet, nous utilisons des techniques de pré-traitement :

syntaxiques : techniques générales telles que la suppression de contenu superflu (*e.g.* les liens **URL**, les balises **HTML** vides, ...).

sémantiques : techniques s'appuyant sur des concepts de **TAL**¹ telles que la lemmatisation basée sur les **POS**² tags.

2.1 Préparation à la tokenisation

Tout d'abord, nous commençons les pré-traitements au niveau d'un document et nous remarquons, en visualisant un échantillon des documents, qu'il s'agit du langage naturel anglais, sans balisage (cf. Figure A.1).

Remplacement des expressions contractées

Nous détectons en premier la présence d'expressions contractées. Or la tokenization qu'on utilisera via la fonction `nltk.tokenize.word_tokenize()` est sensible au guillemet simple³ « ' » désignant les contractions, il faudra

-
1. **Traitement Automatique du Langage Naturel**
 2. **Part-Of-Speech**
 3. et par conséquent aux expressions contractées

ainsi les remplacer par leurs expressions complètes équivalentes. Pour ceci, nous utilisons la fonction `contractions.fix(document)`, encapsulée par la fonction wrapper `replace_contractions(document)` (cf. Figure A.2).

Remarque. Un autre avantage de traiter les expressions contractées est de traiter le cas de la négation, ayant un effet assez conséquent dans le cadre de l’analyse des sentiments : *e.g. I don’t understand* \rightarrow *I do not understand*.

Filtrage des balises HTML auto-fermantes et liens URL

Nous remarquons ensuite la présence de balises **HTML** auto-fermantes⁴ et de liens **URL**, superflus dans le cadre de l’analyse des sentiments (cf. Figure). Nous les filtrons ainsi en utilisant, respectivement, les fonctions `remove_empty_html_tags(document)` et `remove_urls(document)` basées sur des expressions régulières (cf. Figures A.3 et A.4).

Nettoyage syntaxique des phrases

Enfin, nous observons la présence de phrases syntaxiquement mal terminées et/ou mal commencées (cf. Figure), et dont les effets s’avèrent non-favorables lors de la tokenisation. En effet, la tokenisation d’un document les contenant résulte en un token problématique ayant la forme *X.Y* désignant l’un des motifs suivants :

sentence_ends_alphabetic(+)sentence_starts_alphabetic : une phrase qui se termine par une lettre de l’alphabet, suivie d’un ou plusieurs points, suivi(s) d’une lettre de l’alphabet d’une phrase qui commence, sans aucun espace blanc entre la fin de la 1^{re} et le début de la 2^e.

sentence_ends_digit(+)sentence_starts_alphabetic : *idem* que le 1^{er} mais avec une phrase qui se termine par un chiffre et une phrase qui commence par une lettre de l’alphabet.

sentence_ends_alphabetic(+)sentence_starts_digit : *idem* mais avec une phrase qui se termine par une lettre de l’alphabet et une phrase qui commence par un chiffre.

Pour pallier ce problème, nous utilisons une fonction `clean_sentence_anchors(document)` basée sur des expressions régulières et permettant d’obtenir des phrases bien terminées/commencées (cf. Figure A.5).

Remarque. Le motif `digit(+)digit` n’est pas pris en compte, vu que les expressions régulières ne pourrions ainsi le distinguer d’un nombre à virgule.

4. *e.g.* `
`

Ce cas reste ainsi non traité, vu qu'il est plus délicat à contourner, surtout que la présence de nombres en tant qu'amplificateurs d'émotions pourrait être utile pour l'analyse des sentiments, notamment en utilisant les **n-grams** lors de la vectorisation et la sélection des features.

2.2 Tokenisation et normalisation

La deuxième étape de pré-traitement s'effectue au niveau des tokens d'un document.

Normalisation Unicode, encodage ASCII et conversion en minuscule

Afin de n'avoir que des caractères **ASCII** normalisés à traiter, nous effectuons la démarche suivante au sein de la fonction wrapper `remove_non_ascii(tokens)` sur chaque token (cf. Figure A.6) :

1. une normalisation Unicode **NKFD**⁵ via la méthode `unicodedata.normalize('NFKD', token)`.
2. un encodage **ASCII** des caractères normalisés via la méthode `token.encode('ASCII', 'ignore')` en ignorant les caractères non ASCII.
3. un décodage **UTF-8** des octets encodés en **ASCII** via la méthode `token.decode('utf-8', 'ignore')` prenant en compte tous les caractères ASCII.

Une fois les caractères encodés en **ASCII**, nous les convertissons en minuscule en utilisant la fonction `token.lower()`, encapsulée par la fonction wrapper `to_lowercase(tokens)` (cf. Figure A.7).

Remarque. Au cours du projet, nous nous sommes rendu compte que le traitement de motifs contenant des caractères majuscules aurait pu être intéressant dans le cadre de l'analyse des sentiments. Effectivement, étant donné que notre approche primaire était celle du machine learning et non pas celle d'une analyse par lexicons⁶, une approche hybride aurait été encore plus intéressante, mais malheureusement plus chronophage.

Découpage des tokens composés, remplacement des chiffres et filtrage des caractères de ponctuation

Suite à la conversion des tokens en minuscule, nous supprimons les caractères de ponctuation via une fonction `remove_punctuation(tokens)` basée

5. **Normalization Form Compatibility Decomposition**

6. celle-ci nécessitant plus de connaissances en **TAL**

sur des expressions régulières. Cependant, nous nous trouvons ainsi avec des tokens mal formés obtenus par la concaténation de plusieurs tokens. En analysant les résultats, nous trouvons que ces tokens étaient concaténés au préalable par des caractères de ponctuation tels que `_`, `-`, `~`, `...` qui ont été supprimés (cf. Figure A.8). Ainsi nous reportons la suppression des caractères de ponctuation.

Nous commençons alors par découper ces tokens composés, en utilisant la fonction `split_on_character_set(tokens, characters)` basée sur des expressions régulières (cf. Figure A.9).

Après, nous remplaçons les tokens désignant des nombres en chiffres par leurs équivalents en lettres en utilisant la fonction wrapper `replace_numbers(tokens)`, encapsulant la fonction `inflect.engine().number_to_words(token)` (cf. Figure).

Enfin, nous supprimons les caractères de ponctuation en utilisant la fonction susmentionnée `remove_punctuation(tokens)` (cf. Figure A.11).

Remarque. Le coût d'utiliser la fonction de découpage des tokens composés est l'éventuelle perte de sémantique lors de la séparation des mots composés⁷. La fonction de suppression des caractères de ponctuation, quant à elle, supprime la possibilité d'analyser les sentiments à travers des motifs matchant des expressions contenant des caractères tels que «!», «?», voire même les `...`. Toutefois, bien que nous avons sous-estimé les conséquences d'utiliser ces fonctions, les dégâts étaient négligeables par rapport à l'ensemble du dataset. Mais il fallait quand même bien le noter.

Suppression des stopwords et lemmatisation

Après avoir effectué les pré-traitements syntaxiques, nous appliquons des pré-traitements sémantiques s'appuyant sur des notions du **TAL**, notamment le traitement des **stopwords** et la lemmatisation basée sur les **POS** tags.

Pour le traitement des **stopwords**, nous avons considéré le dictionnaire des **stopwords** offert par NLTK pour le langage anglais. Ensuite, nous en avons supprimé le mot «*not*», celui-ci étant utilisé pour le traitement de la négation lors de la vectorisation. Puis, nous y avons ajouté des **stopwords** désignant les mots neutres les plus fréquents⁸ relativement au domaine de la cinéma (*e.g.* *actor* et ses inclinaisons, *movie* et ses synonymes et inclinaisons, `...`). Enfin, la suppression des **stopwords** était effectuée par la fonction wrapper `remove_stopwords(tokens, stopwords)`.

7. *e.g.* *Well-being*, *first-hand*, `...`

8. cf. le fichier `utility_ML.py` pour plus d'informations

La dernière étape de prétraitement consiste en la lemmatisation des tokens en s'appuyant sur leurs catégories grammaticales grâce au **POS-Tagging**. Pour ce faire, nous jugeons, dans le cadre de l'analyse des sentiments, l'importance de considérer les catégories `{noms9, verbes, adjectifs, adverbes}`

Ainsi, pour chaque collection de tokens, les tokens sont taggés par le **POS-Taggeur** collectivement en utilisant la fonction `nltk.pos_tag(tokens)` et ensuite lemmatisés individuellement en considérant leurs tags en utilisant le dictionnaire des catégories à considérer et la fonction `nltk.stem.WordNetLemmatizer().lemmatize(pos_tag)`. L'avantage de cette approche est de convertir un token au lemma le plus sémantiquement cohérent, prenant en compte le contexte du token lemmatisé (cf. Figure A.13).

Une fois la lemmatisation effectuée, nous joignons les tokens normalisés de chaque document afin de pouvoir le vectoriser après.

9. catégorie par défaut

Chapitre 3

Apprentissage du modèle

3.1 Vectorisation

3.2 Feature-Engineering

3.3 Cross validation et calibrage des hyperparamètres

3.4 Sauvegarde du modèle

Chapitre 4

Optimisation

4.1 WordCloud

4.2 Traitement de l'ironie

Chapitre 5

Conclusion

Annexes

Annexe A

Snapshots de pré-traitement

```
In [7]: 1 opinions = df['Avis'][:10]
        2 for opinion in opinions:
        3     print(opinion + "\n")
```

When the young Kevin gets the boat of his dead uncle as a gift, he invites five friends of him to a trip to Catalina Isl and for the weekend. While in the journey, they drink booze, have sex and play games, with each one of them telling his or her greatest fear. Later Kevin drowns in the open sea, the engine stops, and they are haunted and murdered by their greatest innermost fear. Yesterday, my wife, son, daughter and three other friends joined to watch "Haunted Boat" on DVD. With less than 30 minutes running time, the group gave up watching this messy and boring amateurish piece of crap, and we decided to see another film. Later, I decided to watch the rest of this flick to see how bad it could be and it would have been better off going to bed to sleep. The confused story has an awful cinematography and camera work, with a cast that is probably studying to be actors and actresses and in the end this film seems to be a bad project of cinema school. The terrible and pretentious screenplay shows a ridiculous twist in the end, actually a complete mess that made me not understand what the story is all about. Was the girl insane and traveled alone in the boat, imagining the whole situation with imaginary friends? If that is true, are their friend again in the very end fruit of her madness? My vote is one. Title (Brazil): "Viagem Para a Morte" ("Trip to the Death")

Since I first saw Anchors Aweigh in 1945, viewing it on videotape holds a lot of nostalgia for me. At age 15, it was easy for me to be drawn into the first of the great MGM Technicolor musicals. Now I am perhaps most interested in thinking about the future careers of the leading players. Though Sinatra had done a couple of negligible films soon after his emergence after his Dorsey days, as a solo singer, this was his first major film appearance. As another viewer noted, this seems almost to be a warm-up for On the Town. Sinatra may have had to work hard at it, but his dance with Kelly is credible, and he would do better in their next pairings. However, observing his physique, it's easy to see why he was caricatured as a string bean. Who would have imagined that within a decade he would win an academy award for acting, and go on to play many roles as a tough detective or leader in combat. Though Gene Kelly's personality and dancing dominated this film, his winsome performance did not suggest that he would become a major creative force, almost the iconic figure, for MGM musicals, where he developed a style of dance complementary to that of Fred Astaire. Finally, it was strange to see

FIGURE A.1 – échantillon de documents

Before replacing contractions:

=====

When I saw the commercial for this, I was all about seeing it. Now, forgive me, but it's been so long since I've seen it that I don't recall how it went. Suffice it to say, the movie I saw bore no resemblance to the "movie" they sold me on. I was bored, annoyed, and incredibly disappointed by this movie. And if it wasn't bad enough, they had to sink it even further with that awful reggae music. Not exactly mood-setting music for a horror movie, eh mon? I guess if you never saw the commercial (or trailer, I suppose) you may think this is some hot stuff. For my money, the commercial was way better.

After replacing contractions:

=====

When I saw the commercial for this, I was all about seeing it. Now, forgive me, but it is been so long since I have seen it that I do not recall how it went. Suffice it to say, the movie I saw bore no resemblance to the "movie" they sold me on. I was bored, annoyed, and incredibly disappointed by this movie. And if it was not bad enough, they had to sink it even further with that awful reggae music. Not exactly mood-setting music for a horror movie, eh mon? I guess if you never saw the commercial (or trailer, I suppose) you may think this is some hot stuff. For my money, the commercial was way better.

FIGURE A.2 – remplacement des expressions contractées dans un document

Before removing empty HTML tags:
 =====
 Rather like Paul Newman and Steve McQueen with their racing car movies this has all the appearance of a "jollies" project for Robert Redford, as he gets to ski up hill and down dale in the Alpine sunshine.

The story is as light as powdered snow with Redford's small-town boy David Chappellet (what kind of lead name is that?) who with his eyes on the prize of Olympic glory, gets up the nose of, in no particular order, his coach, father and team-mates. Women are a mere side-show in his insular world as evidenced by a fairly distasteful pick-up scene with an old girlfriend in his hometown and then his selfishly petulant pursuit of, heavens above, a free-thinking, independent woman, played by Camilla Sparv. The ski-ing sequences are fine with some good stunt-work involving numerous bumps and scrapes on the piste but their effectiveness is dimmed by our subsequent familiarity with top TV coverage of skiing events down to the present day. Plus I'm not convinced that the Winter Olympics has the same mass identification with the general public as the summer games so that when Redford eventually wins his gold medal in the final reel, I couldn't really be that excited for him one way or another.

Of the actors, Redford, best profile forward, doesn't need to do much and indeed doesn't, while Gene Hackman does better with equally meagre material. Ms Sparv does well as the chief female interest well who treats Redford the way he's doubtless treated every other woman in his chauvinistic way.

In truth though, there's a lack of dramatic tension throughout for which the action sequences don't fully compensate and you don't care a fig for any of the leading characters. One of those films where the actors probably enjoyed making it more than the viewers did watching it....

After removing empty HTML tags:
 =====
 Rather like Paul Newman and Steve McQueen with their racing car movies this has all the appearance of a "jollies" project for Robert Redford, as he gets to ski up hill and down dale in the Alpine sunshine. The story is as light as powdered snow with Redford's small-town boy David Chappellet (what kind of lead name is that?) who with his eyes on the prize of Olympic glory, gets up the nose of, in no particular order, his coach, father and team-mates. Women are a mere side-show in his insular world as evidenced by a fairly distasteful pick-up scene with an old girlfriend in his hometown and then his selfishly petulant pursuit of, heavens above, a free-thinking, independent woman, played by Camilla Sparv. The ski-ing sequences are fine with some good stunt-work involving numerous bumps and scrapes on the piste but their effectiveness is dimmed by our subsequent familiarity with top TV coverage of skiing events down to the present day. Plus I am not convinced that the Winter Olympics has the same mass identification with the general public as the summer games so that when Redford eventually wins his gold medal in the final reel, I could not really be that excited for him one way or another. Of the actors, Redford, best profile forward, does not need to do much and indeed does not, while Gene Hackman does better with equally meagre material. Ms Sparv does well as the chief female interest well who treats Redford the way he is doubtless treated every other woman in his chauvinistic way. In truth though, there is a lack of dramatic tension throughout for which the action sequences do not fully compensate and you do not care a fig for any of the leading characters. One of those films where the actors probably enjoyed making it more than the viewers did watching it....

FIGURE A.3 – suppression des balises auto-fermantes d'un document

Before removing URLs:
 =====
 Did Uwe Boll seriously just rip off the basic idea and dialogue from Se7en?! Why is it so fекking difficult for this douchebag to be original?! He even mentioned in an interview with Gametrailers that he chooses stuff like games to make into movies because the characters, plots, backstories and so on are already there and ready for him to screw with.

Guess it isn't too much of a stretch for him to rip off another movie entirely...

I mean, seriously, what the hell...? Here's something I made in Uwe's 'honor'...

<http://zuucka.deviantart.com/art/Uwe-Boll-is-a-Douchebag-70369862>...

After removing URLs:
 =====
 Did Uwe Boll seriously just rip off the basic idea and dialogue from Se7en?! Why is it so fекking difficult for this douchebag to be original?! He even mentioned in an interview with Gametrailers that he chooses stuff like games to make into movies because the characters, plots, backstories and so on are already there and ready for him to screw with. Guess it is not too much of a stretch for him to rip off another movie entirely... I mean, seriously, what the hell...? Here's something I made in Uwe's 'honor'...

FIGURE A.4 – suppression des URLs d'un document

Before cleaning sentence anchors:
=====

This movie (even calling it a movie is an overstatement) is ridiculously horrible. Normally a huge fan of Eric Roberts in "B" list movies, this tragedy of a flick makes me question his real B list clout! And Charlie, please go back to hoping for a Diagnosis Murder revival rather than this.....you can't blame the nameless eye candy (uhhum...beauty pageant members) for participating in this weak movie, but YOU are a former TV star man! Pull yourself together. Don't even get me started on Stuart Pankin. For the sake of all that is good Stuart, you should have seen this was not necessarily a real movie! Bryan Michael Stoller exemplifies absolute genius only in the fact that he was able to dupe anyone into investing in this picture (money or time).

Really, this was no parody or spoof movie although it tries on a 2nd grade level. Mostly, it is poor writing and acting and camera work and editing and....well poor everything. I watched it because I read an article in some mag about agent MJ's involvement and my interest was peaked due to the lawsuit in which he was involved. I now wonder if the only reason they show him from the shoulders up in the movie is because he, like at the trial, showed up wearing pajama bottoms and barely lucid (wait a second, is he ever really considered lucid?...I digress). And Agent MJ? Is that the best they could come up with for a name for his character? Sheez. What a startling piece of originality! Or, maybe that was supposed to be funny? Putting Marriott into the movie was a nice touch at first, but overdone and annoying after all is said and done.

Spare yourself the grief of watching.....don't say I didn't warn you....

After cleaning sentence anchors:
=====

This movie (even calling it a movie is an overstatement) is ridiculously horrible. Normally a huge fan of Eric Roberts in "B" list movies, this tragedy of a flick makes me question his real B list clout! And Charlie, please go back to hoping for a Diagnosis Murder revival rather than this..... you cannot blame the nameless eye candy (uhhum... beauty pageant members) for participating in this weak movie, but YOU are a former TV star man! Pull yourself together. do not even get me started on Stuart Pankin. For the sake of all that is good Stuart, you should have seen this was not necessarily a real movie! Bryan Michael Stoller exemplifies absolute genius only in the fact that he was able to dupe anyone into investing in this picture (money or time). Really, this was no parody or spoof movie although it tries on a 2nd grade level. Mostly, it is poor writing and acting and camera work and editing and.... well poor everything. I watched it because I read an article in some mag about agent MJ's involvement and my interest was peaked due to the lawsuit in which he was involved. I now wonder if the only reason they show him from the shoulders up in the movie is because he, like at the trial, showed up wearing pajama bottoms and barely lucid (wait a second, is he ever really considered lucid?...I digress). And Agent MJ? Is that the best they could come up with for a name for his character? Sheez. What a startling piece of originality! Or, maybe that was supposed to be funny? Putting Marriott into the movie was a nice touch at first, but overdone and annoying after all is said and done. Spare yourself the grief of watching..... do not say I did not warn you....

FIGURE A.5 – nettoyage syntaxique des phrases mal commencées/terminées d'un document

Before Unicode normalization and ASCII encoding:
=====

In the early 00's, production companies had a short-lived craze for supernatural genre movies in France after "The Crimson Rivers" and "Brotherhood of the Wolf" turned out to be hits, so several movies were green-lit or saved from their "direct-to-video" fate. However, France, as opposed to the US, UK or Italy, has little tradition of fantasy B-movies and it turned out quickly that "Samourais", "Bloody Mallory" or the "Crimson Rivers" sequel were ill-advised attempts at recreating a kind of magic that had never existed in French cinema in the first place. As they flopped, producers have gone back to their usual fare: derivative farces or the umpteenth self-referential tribute to French New Wave by a former critic from "Les Cahiers du cinéma".

"Brocéliande" could only have been green-lit during this short window, as it serves no other discernible purpose. It's your by-the-book

After Unicode normalization and ASCII encoding:
=====

['In', 'the', 'early', '00', 's', ',', 'production', 'companies', 'had', 'a', 'short-lived', 'craze', 'for', 'supernatural', 'genre', 'movies', 'in', 'France', 'after', '...', 'The', 'Crimson', 'Rivers', '...', 'and', '...', 'Brotherhood', 'of', 'the', 'Wolf', '...', 'turned', 'out', 'to', 'be', 'hits', '...', 'so', 'several', 'movies', 'were', 'green-lit', 'or', 'saved', 'from', 'their', '...', 'direct-to-video', '...', 'fate', '...', 'However', '...', 'France', '...', 'as', 'opposed', 'to', 'the', 'US', '...', 'UK', 'or', 'Italy', '...', 'has', 'little', 'tradition', 'of', 'fantasy', 'B-movies', 'and', 'it', 'turned', 'out', 'quickly', 'that', '...', 'Samourais', '...', '...', 'Bloody', 'Mallory', '...', 'or', 'the', '...', 'Crimson', 'Rivers', '...', 'sequel', 'were', 'ill-advised', 'attempts', 'at', 'recreating', 'a', 'kind', 'of', 'magic', 'that', 'had', 'never', 'existed', 'in', 'French', 'cinema', 'in', 'the', 'first', 'place', '...', 'As', 'they', 'flopped', '...', 'producers', 'have', 'gone', 'back', 'to', 'their', 'usual', 'fare', '...', 'derivative', 'farces', 'or', 'the', 'umpteenth', 'self-referential', 'tribute', 'to', 'French', 'New', 'Wave', 'by', 'a', 'former', 'critic', 'from', 'the', 'Les', 'Cahiers', 'du', 'cinéma', '...', '...', '...', 'Brocéliande', '...', 'could', 'only', 'have', 'been', 'green-lit', 'during', 'this', 'short', 'window', '...', 'as', 'it', 'serves', 'no', 'other', 'discernible', 'purpose', '...', 'it', 'is', 'your', 'by-the-book']

FIGURE A.6 – normalisation Unicode, et encodage ASCII des tokens d'un document

```

Before conversion of characters to lower case:
=====
I never trust the opinions of anyone regarding a film. That goes for critics as well. Sure, if it gets positive reviews that's OK and a plus, but most films that get critical rave I hate. I enjoyed this film for what it was, an entertaining film. It takes you out of your life for a couple hours and into a fictional character... that being Catherine Trammell. Sharon Stone is awesome in this role, just like she was in the first one. Anyone who says she is horrible in this film must have felt the same in the first one b/c she is back acting the same way she did in Basic Instinct 1. Catherine is hers and she plays her to perfection. Her one liners are great, much like in the first one. Who can forget in the first film when she tells the cops, "If you're gonna arrest me do it...otherwise get the f**k out of here!" Great scene, and believe me, she does it again in this one. I was captivated by her. Her outfits, the way she smoked her cigarettes, believe me, its worth the price just to see Stone's performance. I cannot wait for this film to be released on DVD, uncut, because I can only imagine how much better it is going to be. And yes, there are lots of twists, as in the first one, including the ending!

After conversion of characters to lower case:
=====
['i', 'never', 'trust', 'the', 'opinions', 'of', 'anyone', 'regarding', 'a', 'film', '.', 'that', 'goes', 'for', 'critics', 'as', 'well', '.', 'sure', '.', 'if', 'it', 'gets', 'positive', 'reviews', 'that', 'is', 'ok', 'and', 'a', 'plus', '.', 'but', 'most', 'films', 'that', 'get', 'critical', 'rave', 'i', 'hate', '.', 'i', 'enjoyed', 'this', 'film', 'for', 'what', 'it', 'was', '.', 'an', 'entertaining', 'film', '.', 'it', 'takes', 'you', 'out', 'of', 'your', 'life', 'for', 'a', 'couple', 'hours', 'and', 'into', 'a', 'fictional', 'character', '.', 'that', 'being', 'catherine', 'trammell', '.', 'sharon', 'stone', 'is', 'awesome', 'in', 'this', 'role', '.', 'just', 'like', 'she', 'was', 'in', 'the', 'first', 'one', '.', 'anyone', 'who', 'says', 'she', 'is', 'horrible', 'in', 'this', 'film', 'must', 'have', 'felt', 'the', 'same', 'in', 'the', 'first', 'one', 'b/c', 'she', 'is', 'back', 'acting', 'the', 'same', 'way', 'she', 'did', 'in', 'basic', 'instinct', '1', '.', 'catherine', 'is', 'hers', 'and', 'she', 'plays', 'her', 'to', 'perfection', '.', 'her', 'one', 'liners', 'are', 'great', '.', 'much', 'like', 'in', 'the', 'first', 'one', '.', 'who', 'can', 'forget', 'in', 'the', 'first', 'film', 'when', 'she', 'tells', 'the', 'cops', '.', 'if', 'you', 'are', 'going', 'to', 'arrest', 'me', 'do', 'it', '...', 'otherwise', 'get', 'the', 'f**k', 'out', 'of', 'here', '!', 'great', 'scene', '.', 'and', 'believe', 'me', 'she', 'does', 'it', 'again', 'in', 'this', 'one', '.', 'i', 'was', 'captivated', 'by', 'her', '.', 'her', 'outfits', 'the', 'way', 'she', 'smoked', 'her', 'cigarettes', 'believe', 'me', 'its', 'worth', 'the', 'price', 'just', 'to', 'see', 'stone', 's', 'performance', '.', 'i', 'can', 'not', 'wait', 'for', 'this', 'film', 'to', 'be', 'released', 'on', 'dvd', 'uncut', 'because', 'i', 'can', 'only', 'imagine', 'how', 'much', 'better', 'it', 'is', 'going', 'to', 'be', 'and', 'yes', 'there', 'are', 'lots', 'of', 'twists', 'as', 'in', 'the', 'first', 'one', 'including', 'the', 'ending', '!']

```

FIGURE A.7 – conversion en minuscule des tokens d'un document

```

In [88]: 1 sentence = "this is a compounded-token, this is another compounded/token and this is yet another compounded-token"
          2
          3 print(remove_punctuation(word_tokenize(sentence)))

['this', 'is', 'a', 'compoundedtoken', 'this', 'is', 'another', 'compoundedtoken', 'and', 'this', 'is', 'yet', 'another', 'compoundedtoken']

```

FIGURE A.8 – suppression des caractères de ponctuation, avant le découpage de tokens composés

```

In [90]: 1 sentence = "this is a compounded-token, this is another compounded/token and this is yet another compounded-token"
          2
          3 print(split_on_characterset(word_tokenize(sentence), r'[/\~_-]'))

['this', 'is', 'a', 'compounded', 'token', 'this', 'is', 'another', 'compounded', 'token', 'and', 'this', 'is', 'yet', 'another', 'compounded', 'token']

```

FIGURE A.9 – découpage des tokens composés

Before replacing digits with letters:
=====

This movie is like Happiness meets Lost in Translation with a Sixth Sense ending (or maybe a Crying Game surprise), and the best soundtrack I've probably ever heard...if that all make sense.

The first 30 seconds pretty much tells you you're in for a twisted ride. (I was surprised no one walked out right away during the Brooklyn premiere.) But from there, the film settles down into a talk-fest between two really damaged people, Daphne and Buddy.

They're lonely, mess-up, and boy do they talk about sex. Daphne brings to life her most interesting tales of escorting, some are quite funny (Mr. Chang) some disturbing (the Harlan scenes with music that tells us what we see might now be what's going on, or what Daphne is really feeling), and because I have a friend who used to escort, I might add, most seem quite real.

You Are Alone is multi-layered and mostly brilliant. Okay, maybe a couple minutes less of the talking, and I don't know that we'd have missed anything.

Then again, I need to see it again knowing the ending.

I like this movie.

(The director asked people in the Brooklyn audience to write a review on IMDb because a lot of people read them. Request granted.)

Before replacing digits with letters:
=====

['this', 'movie', 'is', 'like', 'happiness', 'meets', 'lost', 'in', 'translation', 'with', 'a', 'sixth', 'sense', 'ending', '(', 'or', 'maybe', 'a', 'crying', 'game', 'surprise', ')', ' ', 'and', 'the', 'best', 'soundtrack', 'i', 'have', 'probably', 'ever', 'heard', 'if', 'that', 'all', 'make', 'sense', ' ', 'the', 'first', 'thirty', 'seconds', 'pretty', 'much', 'tells', 'you', 'you', 'are', 'in', 'for', 'a', 'twisted', 'ride', ' ', '(', 'i', 'was', 'surprised', 'no', 'one', 'walked', 'out', 'right', 'away', 'during', 'the', 'brooklyn', 'premiere', ' ', 'but', 'from', 'there', 'the', 'film', 'settles', 'down', 'into', 'a', 'talk', 'fest', 'between', 'two', 'really', 'damaged', 'people', ' ', 'daphne', 'and', 'buddy', ' ', 'they', 'are', 'lonely', ' ', 'mess', 'up', ' ', 'and', 'boy', 'do', 'they', 'talk', 'about', 'sex', ' ', 'daphne', 'brings', 'to', 'life', 'her', 'most', 'interesting', 'tales', 'of', 'escorting', ' ', 'some', 'are', 'quite', 'funny', '(', 'mr.', 'chang', ')', ' ', 'some', 'disturbing', '(', 'the', 'harlan', 'scenes', 'with', 'music', 'that', 'tells', 'us', 'what', 'we', 'see', 'might', 'now', 'be', 'what', 'is', 'going', 'on', ' ', 'or', 'what', 'daphne', 'is', 'really', 'feeling', ')', ' ', 'and', 'because', 'i', 'have', 'a', 'friend', 'who', 'used', 'to', 'escort', ' ', 'i', 'might', 'add', ' ', 'most', 'seem', 'quite', 'real', ' ', 'you', 'are', 'alone', 'is', 'multi', 'layered', 'and', 'mostly', 'brilliant', ' ', 'okay', ' ', 'maybe', 'a', 'couple', 'minutes', 'less', 'of', 'the', 'talking', ' ', 'and', 'i', 'do', 'not', 'know', 'that', 'we', 'would', 'have', 'missed', 'anything', ' ', 'then', 'again', ' ', 'i', 'need', 'to', 'see', 'it', 'again', 'knowing', 'the', 'ending', ' ', 'i', 'like', 'this', 'movie', ' ', '(', 'the', 'director', 'asked', 'people', 'in', 'the', 'brooklyn', 'audience', 'to', 'write', 'a', 'review', 'on', 'imdb', 'because', 'a', 'lot', 'of', 'people', 'read', 'them', ' ', 'request', 'granted', ' ', ')']

FIGURE A.10 – remplacement des tokens désignant des chiffres par leurs équivalents en lettres

Before removing punctuation:
=====

This movie is like Happiness meets Lost in Translation with a Sixth Sense ending (or maybe a Crying Game surprise), and the best soundtrack I've probably ever heard...if that all make sense.

The first 30 seconds pretty much tells you you're in for a twisted ride. (I was surprised no one walked out right away during the Brooklyn premiere.) But from there, the film settles down into a talk-fest between two really damaged people, Daphne and Buddy.

They're lonely, mess-up, and boy do they talk about sex. Daphne brings to life her most interesting tales of escorting, some are quite funny (Mr. Chang) some disturbing (the Harlan scenes with music that tells us what we see might now be what's going on, or what Daphne is really feeling), and because I have a friend who used to escort, I might add, most seem quite real.

You Are Alone is multi-layered and mostly brilliant. Okay, maybe a couple minutes less of the talking, and I don't know that we'd have missed anything.

Then again, I need to see it again knowing the ending.

I like this movie.

(The director asked people in the Brooklyn audience to write a review on IMDb because a lot of people read them. Request granted.)

After removing punctuation:
=====

['this', 'movie', 'is', 'like', 'happiness', 'meets', 'lost', 'in', 'translation', 'with', 'a', 'sixth', 'sense', 'ending', 'or', 'maybe', 'a', 'crying', 'game', 'surprise', 'and', 'the', 'best', 'soundtrack', 'i', 'have', 'probably', 'ever', 'heard', 'if', 'that', 'all', 'make', 'sense', 'the', 'first', 'thirty', 'seconds', 'pretty', 'much', 'tells', 'you', 'you', 'are', 'in', 'for', 'a', 'twisted', 'ride', 'i', 'was', 'surprised', 'no', 'one', 'walked', 'out', 'right', 'away', 'during', 'the', 'brooklyn', 'premiere', 'but', 'from', 'there', 'the', 'film', 'settles', 'down', 'into', 'a', 'talk', 'fest', 'between', 'two', 'really', 'damaged', 'people', 'daphne', 'and', 'buddy', 'they', 'are', 'lonely', 'mess', 'up', 'and', 'boy', 'do', 'they', 'talk', 'about', 'sex', 'daphne', 'brings', 'to', 'life', 'her', 'most', 'interesting', 'tales', 'of', 'escorting', 'some', 'are', 'quite', 'funny', 'mr', 'chang', ' ', 'some', 'disturbing', 'the', 'harlan', 'scenes', 'with', 'music', 'that', 'tells', 'us', 'what', 'we', 'see', 'might', 'now', 'be', 'what', 'is', 'going', 'on', 'or', 'what', 'daphne', 'is', 'really', 'feeling', 'and', 'because', 'i', 'have', 'a', 'friend', 'who', 'used', 'to', 'escort', 'i', 'might', 'add', 'most', 'seem', 'quite', 'real', 'you', 'are', 'alone', 'is', 'multi', 'layered', 'and', 'mostly', 'brilliant', 'okay', 'maybe', 'a', 'couple', 'minutes', 'less', 'of', 'the', 'talking', 'and', 'i', 'do', 'not', 'know', 'that', 'we', 'would', 'have', 'missed', 'anything', 'then', 'again', 'i', 'need', 'to', 'see', 'it', 'again', 'knowing', 'the', 'ending', 'i', 'like', 'this', 'movie', 'the', 'director', 'asked', 'people', 'in', 'the', 'brooklyn', 'audience', 'to', 'write', 'a', 'review', 'on', 'imdb', 'because', 'a', 'lot', 'of', 'people', 'read', 'them', 'request', 'granted']

FIGURE A.11 – suppression des caractères de ponctuation

```

Before removing stopwords:
=====
This movie is like Happiness meets Lost in Translation with a Sixth Sense ending (or maybe a Crying Game surprise), and the best soundtrack I've probably ever heard...if that all make sense.<br /><br />The first 30 seconds pretty much tells you you're in for a twisted ride. (I was surprised no one walked out right away during the Brooklyn premiere.) But from there, the film settles down into a talk-fest between two really damaged people, Daphne and Buddy.<br /><br />They're lonely, mess-up, and boy do they talk about sex. Daphne brings to life her most interesting tales of escorting, some are quite funny (Mr. Chang) some disturbing (the Harlan scenes with music that tells us what we see might now be what's going on, or what Daphne is really feeling), and because I have a friend who used to escort, I might add, most seem quite real. <br /><br />You Are Alone is multi-layered and mostly brilliant. Okay, maybe a couple minutes less of the talking, and I don't know that we'd have missed anything. <br /><br />Then again, I need to see it again knowing the ending.<br /><br />I like this movie.<br /><br />(The director asked people in the Brooklyn audience to write a review on IMDb because a lot of people read them. Request granted.)

After removing stopwords:
=====
['like', 'happiness', 'meets', 'lost', 'translation', 'sixth', 'sense', 'ending', 'maybe', 'crying', 'game', 'surprise', 'best', 'soundtrack', 'probably', 'ever', 'heard', 'make', 'sense', 'first', 'thirty', 'seconds', 'pretty', 'much', 'tells', 'twisted', 'ride', 'surprised', 'no', 'one', 'walked', 'right', 'away', 'brooklyn', 'premiere', 'settles', 'talk', 'fest', 'two', 'really', 'damaged', 'people', 'daphne', 'buddy', 'lonely', 'mess', 'boy', 'talk', 'sex', 'daphne', 'brings', 'life', 'interesting', 'tales', 'escorting', 'quite', 'funny', 'mr', 'chang', 'disturbing', 'harlan', 'music', 'tells', 'us', 'see', 'might', 'going', 'daphne', 'really', 'feeling', 'friend', 'used', 'escort', 'might', 'add', 'seem', 'quite', 'real', 'alone', 'multi', 'layered', 'mostly', 'brilliant', 'okay', 'maybe', 'couple', 'minutes', 'less', 'talking', 'not', 'know', 'would', 'missed', 'anything', 'need', 'see', 'knowing', 'ending', 'like', 'asked', 'people', 'brooklyn', 'audience', 'write', 'review', 'imdb', 'lot', 'people', 'read', 'request', 'granted']

```

FIGURE A.12 – suppression des stopwords

```

Sentence: I love this food
POS TAGGING : [('I', 'PRP'), ('love', 'VBP'), ('this', 'DT'), ('food', 'NN')]
Lemmatization without POS TAGGING nor normalization: ['I', 'love', 'this', 'food']
Lemmatization after normalization: ['love', 'food']

Sentence: I'm in love with eating this food
POS TAGGING : [('I', 'PRP'), ('m', 'VBP'), ('in', 'IN'), ('love', 'NN'), ('with', 'IN'), ('eating', 'VBG'), ('this', 'DT'), ('food', 'NN')]
Lemmatization without POS TAGGING nor normalization: ['I', 'm', 'in', 'love', 'with', 'eating', 'this', 'food']
Lemmatization after normalization: ['love', 'eat', 'food']

Sentence: I will eat this food lovingly
POS TAGGING : [('I', 'PRP'), ('will', 'MD'), ('eat', 'VB'), ('this', 'DT'), ('food', 'NN'), ('lovingly', 'RB')]
Lemmatization without POS TAGGING nor normalization: ['I', 'will', 'eat', 'this', 'food', 'lovingly']
Lemmatization after normalization: ['eat', 'food', 'lovingly']

Sentence: I find this food lovable
POS TAGGING : [('I', 'PRP'), ('find', 'VBP'), ('this', 'DT'), ('food', 'NN'), ('lovable', 'JJ')]
Lemmatization without POS TAGGING nor normalization: ['I', 'find', 'this', 'food', 'lovable']
Lemmatization after normalization: ['find', 'food', 'lovable']

```

FIGURE A.13 – lemmatization en utilisant le POS-Tagging