

Si l'algorithmique m'était contée...

Comment se servir des algorithmes pour faire de la détection de langue ?

Un Apprentissage Par Problème (APP) destiné aux étudiants
du module Algorithmique avancée 1

Mars 2022



Présentation du problème

Vous faites partie d'une équipe de recherche en linguistique informatique. Ce champ interdisciplinaire met en avant des approches statistiques pour l'étude du langage naturel.

Votre tâche est de caractériser quantitativement différentes langues européennes par leurs fréquences d'utilisation des lettres de l'alphabet (les alphabets peuvent être différents d'une langue à l'autre mais elles ont toutes en commun les lettres latines).

Vous avez à votre disposition un corpus de texte classique dans 7 langues provenant du Project Gutenberg (<https://www.gutenberg.org/>) :

- Français : "Les fables de La Fontaine" de Jean de la Fontaine ; "La comédie humaine" de Honoré de Balzac
- Anglais : "Richard III", "Hamlet" de William Shakespeare; "Moby-Dick; or The Whale" de Herman Melville
- Portugais : "Os Lusíadas" de Luís Vaz de Camões; "Os Maias" de José Maria Eça de Queirós
- Espagnol : "Don Quijote" de Miguel de Cervantes Saavedra
- Allemand : "Faust Der Tragödie erster Teil" de Johann Wolfgang von Goethe
- Italien : "La Divina Commedia di Dante" de Dante Alighieri
- Néerlandais : "De ondergang der Eerste Wared" de Willem Bilderdijk

Une méthode naïve de caractérisation d'une langue peut être de comparer la lettre la plus fréquente ; la seconde lettre la plus fréquente et ainsi de suite jusqu'à trouver une différence entre deux langues.

Cette méthode n'est certainement pas parfaite mais vous décidez de l'utiliser pour vérifier des observations plus générales de classification des langues européennes :

En effet ces langues peuvent être classées en deux groupes :

(1) Langues romanes :

- |_ français
- |_ italien
- |_ espagnol
- |_ portugais

(2) Langues germaniques :

- |_ anglais
- |_ allemand
- |_ néerlandais

La question se pose de différencier d'avantage ces langues au sein de ces groupes de langues.

Il se peut que le corpus de texte ne soit pas suffisant pour avoir une représentation exacte de la langue. La taille de l'échantillon ou la date de parution (par exemple : Le texte des "Fables de La Fontaine" est en « vieux français ») peuvent influencer la distribution.

Attention à vérifier les données, il ne faudrait pas parasiter la caractérisation avec des mots dans une autre langue...

Quelques semaines plus tard, l'un de vos collègues vous présente deux textes en galicien : "Follas Novas" et "Cantares Gallegos" d'une poétesse nommée Rosalía de Castro.

*Adiós ríos, adiós fontes,
Adiós regatos pequenos,
Adiós vista dos meus ollos,
Non sei cándido nos veremos.*

...

Rosalía de Castro

Intrigué par la langue qu'il ne sait être ni complètement du portugais ni complètement de l'espagnol, Il vous demande si votre méthode de caractérisation permet de dire si le galicien est plus proche du portugais ou de l'espagnol...

Ressources pour traiter la situation problème

Tri comptage (ou tri par dénombrement) :

https://fr.wikipedia.org/wiki/Tri_comptage

https://haltode.fr/algo/tri/tri_denombrement.html

<http://www.dailly.info/Tri-par-casier>

Histogramme:

<https://fr.wikipedia.org/wiki/Histogramme>

Ressources Moodle pour traiter la situation-problème

[Toolbook - Tri par comptage - Les images dans python](#)

Calendrier du traitement de l'APP :

Timing séance « aller » et travail individuel

Phases et Etapes		Tâches
Phases A séance « ALLER »	1 10min	Organiser l'équipe : <ul style="list-style-type: none"> Se répartir les fonctions indispensables (voir page 11) <i>Le barreur prend connaissance des étapes à parcourir et garde le cap</i> <i>Le gardien du temps s'engage à surveiller le timing</i>
	2 10min	Prendre connaissance du document fourni : <ul style="list-style-type: none"> Chacun effectue un premier survol du cahier pour se familiariser avec le contenu
	3 10min	Comprendre et clarifier le problème : à partir de la p. 3 : <ul style="list-style-type: none"> Quel est au juste le problème que nous allons traiter ? <i>Le scribe commence à noter ce qui apparaît dans les échanges (mots-clés, concepts, idées, ...)</i>
	4 30min	Etablir ensemble des pistes pour traiter le problème : <ul style="list-style-type: none"> Etablir une liste de questions pertinentes auxquelles il faudra répondre Faire le point sur ce que l'équipe connaît (et ne connaît pas) Le cas échéant, établir une liste de simplifications, de restrictions en vue de limiter la portée du problème (si nécessaire, voir avec le tuteur) Etablir une liste des productions attendues Envisager différentes pistes pour avancer dans le traitement <i>L'activateur lance et relance la discussion quand c'est nécessaire</i>
	5 20min	Préciser les acquis d'apprentissage : <ul style="list-style-type: none"> Que faut-il (ré-)apprendre / découvrir pour traiter le problème ? A quelles questions chacun de nous devra-t-il être capable de répondre à la fin de la séance « RETOUR » ? Que faudra-t-il être capable de faire ?
	6 15min	Etablir un plan d'action : <ul style="list-style-type: none"> Déterminer les informations à recueillir pour confirmer ou invalider les pistes énumérées Dresser la liste des tâches à accomplir et des livrables à préparer par chacun <u>avant</u> la prochaine séance, ... <i>Le secrétaire note ce qui est décidé et s'arrange pour le communiquer aux autres membres de l'équipe</i>

Phase B Travail Indiv.	7 de 6 à 10 h	Travail individuel : <ul style="list-style-type: none"> Mettre en œuvre le plan d'action établi à l'étape 6 : chacun effectue le travail décidé et prépare ce qu'il va apporter à la séance « Retour »
---------------------------	------------------	---

Phases C Séance « RETOUR »		(détails p. 9)
-------------------------------	--	----------------

Timing séance « retour »

Etape	Timing indicatif	Activités	Productions possibles / Produits attendus	Pistes possibles pour le tuteur
8	10 min	Organiser le groupe	<ul style="list-style-type: none"> ▪ Une (re)distribution des fonctions ▪ Une organisation de l'équipe ▪ Un planning de l'étape 9 ▪ Préciser la production attendue 	<ul style="list-style-type: none"> ▪ Demander s'il est utile de changer les attributions des fonctions ▪ Insister sur la planification de l'étape 9 : quelles sont les différentes tâches à accomplir et combien de temps y consacrer ? ▪ Suggérer de préparer un ou deux posters avec les éléments clés de la réponse de l'équipe.
9	60 min	Valider les apprentissages, les solutions, les livrables	<ul style="list-style-type: none"> ▪ Les réponses aux questions formulées lors de la séance « aller » ▪ L'inventaire de ce qui a été compris et appris ▪ L'inventaire de ce qui reste à approfondir ▪ La synthèse à présenter aux autres équipes 	<ul style="list-style-type: none"> ▪ Confrontation sur les points délicats ▪ Comparer les apports de chacun ▪ Identifier les différences, les complémentarités ▪ Mettre l'accent sur la production d'une synthèse : qu'aurait-on envie de communiquer aux autres ?

Etape	Timing indicatif	Activités	Productions possibles / Produits attendus	Pistes possibles pour le tuteur
10	25 min	QCM	<ul style="list-style-type: none"> ▪ Répondre aux QCM 	<ul style="list-style-type: none"> ▪ Un par table, sans document
11	25 min	Correction QCM	<ul style="list-style-type: none"> ▪ 	<ul style="list-style-type: none"> ▪ Faire un diaporama des réponses

Des fonctions pour faciliter le travail en équipe...






Pour que le travail en équipe se déroule bien et qu'il soit efficace, un peu d'**organisation** est nécessaire... Le tuteur vous aura remis des fiches/cartes qui décrivent différentes fonctions à assumer pour atteindre cet objectif.

Le verso de chaque carte précise en quoi consiste la fonction définie par la carte. Examinez les cartes et répartissez les fonctions entre les membres. Chacun dispose devant lui (ou elle !) la/les carte(s) qui lui est/sont attribuée(s) de façon à ce que chaque membre puisse voir qui prend en charge quelle(s) fonction(s).




Parmi les fonctions proposées, la fonction « **Participant actif** » doit être assumée par chacun des membres !

Quelques fonctions à répartir :

Fonctions indispensables :

Barreur		Vous veillez à l'avancement du travail. Vous faites en sorte que l'équipe suive les étapes imposées ou qu'elle a décidé de suivre. Vous évitez que l'équipe se fourvoie, perde du temps dans des pistes sans issue.
Activateur		Vous amenez chaque membre de l'équipe à contribuer activement aux travaux ; vous n'oubliez ni le scribe, ni le secrétaire ! En cas de nécessité de répartition de tâches, vous veillez à ce que chaque membre contribue de manière équitable.
Gardien du temps		Vous veillez à la bonne utilisation du temps disponible. Vous attirez l'attention sur le risque de prendre du retard.
Scribe		Sur l'espace de travail commun (p. ex. : flip chart), vous notez les idées importantes, les questions en suspens, les schémas qui émergent lors des discussions, mais sans imposer vos propres points de vue. Vous gérez les feuilles du flip chart pour que l'information utile soit visible pour tous les membres de l'équipe. Vous n'oubliez pas de participer aux discussions !
Secrétaire		Vous produisez une synthèse des éléments importants issus des discussions : ceux qu'il faut conserver pour la suite du travail. Vous consignez toutes les informations nécessaires à la poursuite du travail : les décisions prises, les échéances déterminées, les prochains rendez-vous, les plans de travail collectifs et/ou individuels, etc. Vous diffusez vos productions et les autres documents nécessaires à l'ensemble des membres de l'équipe. Vous n'oubliez pas de participer aux discussions !

Fonctions pouvant être utiles :

Circulateur de parole		Vous faites en sorte que chaque membre de l'équipe puisse s'exprimer. Vous incitez les membres en retrait à prendre la parole ; vous n'oubliez ni le scribe, ni le secrétaire ! Vous empêchez l'un ou l'autre membre de l'équipe de mobiliser la parole au détriment des autres.
Porte-parole		Vous présentez l'état ou les résultats du travail de votre équipe d'une manière synthétique et complète, sans marquer de préférence pour votre propre point de vue. Vous utilisez tous les moyens nécessaires pour une communication efficace.
Faiseur de point		Vous faites périodiquement le point sur l'état d'avancement : où en est l'équipe ? qu'est-ce qui est fait ? qu'est-ce qui reste à faire ? que savons-nous et que ne savons-nous pas ? Vous aidez le scribe à noter ces éléments sur l'espace de travail commun.
...		<i>Le cas échéant, ajoutez une fonction qui vous semble utile ou nécessaire</i>