Problématiques liées à l'élaboration d'un corpus de chansons situées (entre autres) temporellement

Gaël Lejeune (gael.lejeune@sorbonne-universite.fr)

21 avril 2022

STIH, Sorbonne University

Pourquoi des chansons ? et Comment ?

Par intérêt pour la variation

Pourquoi des chansons ? et Comment ?

Par intérêt pour la variation

Pourquoi des chansons ? et Comment ?

Par intérêt pour la variation

Double intérêt

Linguistiquement

- Un genre textuel peu étudié
- Un intérêt socio-linguistique évident
- Des états de langue très intéressant diachroniquement
- Un questionnement par rapport aux observables du TAL (mots et phrases en particulier)

Pourquoi des chansons ? et Comment ?

Par intérêt pour la variation

Double intérêt

Linguistiquement

- Un genre textuel peu étudié
- Un intérêt socio-linguistique évident
- Des états de langue très intéressant diachroniquement
- Un questionnement par rapport aux observables du TAL (mots et phrases en particulier)

Informatiquement

- Un intérêt pédagogique
- Des problématiques de scraping
- Des questions de dé-duplication
- Un terrain de jeu original

Corpus de chansons en français/ aux artistes français ou francophones

Outline

Collecte des données

Collecte des données

Comment faire un corpus ?

Pas de corpus de référence disponible, ou alors inutilisable (*million song corpus*)

Comment faire un corpus ?

Pas de corpus de référence disponible, ou alors inutilisable (*million song corpus*)

Alors il faut constituer en respectant autant que possible les critères suivants [**?**] :

signifiance : cohérence et pertinence par rapport à l'**objet étudié**

acceptabilité représentativité, régularité et complétude vis-à-vis de la **finesse et la richesse d'analyse** visées

exploitabilité homogénéité et volume suffisants pour la **significativité statistique**

Comment faire un corpus ?

Pas de corpus de référence disponible, ou alors inutilisable (*million song corpus*)

Alors il faut constituer en respectant autant que possible les critères suivants [**?**] :

signifiance : cohérence et pertinence par rapport à l'**objet étudié**

acceptabilité représentativité, régularité et complétude vis-à-vis de la **finesse et la richesse d'analyse** visées

exploitabilité homogénéité et volume suffisants pour la **significativité statistique**

Let's go !

API genius (I)

Genius

- Des données en grand nombre (millions d'entrées)
- En évolution permanente (plateforme collaborative)
- API facile à utiliser (ce qui ne gâche rien)

Genius API

- Requêtage en ligne de commande ou en Python pur (package `lyricsgenius`)
- Pas de grosse contrainte pour obtenir la clé (*contrary to* Twitter)
- Possibilité de requêter par artiste ou par mots-clés
- Beaucoup de méta-données

API Genius (II)

Pour obtenir une liste d'artistes, utilisons Wikipédia :

- Chanteur Français XXème (887 entrées)
- Chanteuse Française XXème (690 entrées)
- Chanteur Français XXIème (838 entrées)
- Chanteuse Français XXIème (650 entrées)

Complété avec 467 groupes (avec le même déséquilibre temporel)

Collecte à surveiller comme le lait sur le feu (ambiguïtés, déconnexions . . .) avec 529 ≪ négatifs (dont Barbara par exemple) et probablement des vrais négatifs (en particulier chez les ≪ anciens ≫) et des faux positifs non repérés (malgré un filtrage manuel)

Repérons donc déjà un problème de représentativité, et de complétude

API Genius (II)

Collecte 1

1254 artistes ont amené un résultat (529 négatifs pour rappel)

Caractéristique Artistes	Présente	Absente
Description	309 (25%, 843.70 car.)	942
Chansons (méta données)	974 (78% 35.75/artiste)	277
Chansons (méta + paroles)	867 (69% 20.51/artiste)	384
Caractéristique Chansons	**Présente**	**Absente**
Avec paroles	17784 (51.08%)	17035 (48.92%)
Avec date	13867 (39.83%)	20952 (60.17%)
Du XXième	2142 (15.45%)	11725 (84.55%)

Table 1 – Caractéristiques de la première collecte

Croisement avec Paroles.net

Compléter les trous dans la raquette

- Ici c'est le contraire : pas d'API mais une meilleure représentation linguistique et diachronique
- Crawling focalisé à partir de la page d'accueil de `paroles.net`
- Réalisé avec Trafilatura[**?**]
- scraping avec BeautifulSoup (enfin plus ou moins)

83.896 fichiers récupérés, 83.138 avec paroles (VS 17.784 chez Genius)

Croisement avec Paroles.net

Compléter les trous dans la raquette

- Ici c'est le contraire : pas d'API mais une meilleure représentation linguistique et diachronique
- Crawling focalisé à partir de la page d'accueil de `paroles.net`
- Réalisé avec Trafilatura[**?**]
- scraping avec BeautifulSoup (enfin plus ou moins)

83.896 fichiers récupérés, 83.138 avec paroles (VS 17.784 chez Genius)

Et le problème reste entier

- 6728 auteurs (VS 1251)
- 70779 titres (VS 35201)
- 339 artistes en commun, 8608 titres en commun

Confrontation et exploitation des données(I)

Quelques défis

- Compléter les paroles manquantes

Confrontation et exploitation des données(I)

Quelques défis

- Compléter les paroles manquantes
- Opérer un filtrage linguistique ?

Confrontation et exploitation des données(I)

Quelques défis

- Compléter les paroles manquantes
- Opérer un filtrage linguistique ?
- Ajouter des paroles au répertoire d'un artiste

Confrontation et exploitation des données(I)

Quelques défis

- Compléter les paroles manquantes
- Opérer un filtrage linguistique ?
- Ajouter des paroles au répertoire d'un artiste
- Choisir/fabriquer la meilleure "version"

Confrontation et exploitation des données(I)

Quelques défis

- Compléter les paroles manquantes
- Opérer un filtrage linguistique ?
- Ajouter des paroles au répertoire d'un artiste
- Choisir/fabriquer la meilleure "version"
- Gérer le déséquilibre des méta-données

Confrontation et exploitation des données(I)

Quelques défis

- Compléter les paroles manquantes
- Opérer un filtrage linguistique ?
- Ajouter des paroles au répertoire d'un artiste
- Choisir/fabriquer la meilleure "version"
- Gérer le déséquilibre des méta-données
- Croiser avec des bases de connaissance

Confrontation et exploitation des données(II)

Tirer (enfin !) tirer partie du corpus

- De meilleurs corpus pour mes étudiants déjà !

Confrontation et exploitation des données(II)

Tirer (enfin !) tirer partie du corpus

- De meilleurs corpus pour mes étudiants déjà !
- Étude des noms de lieux
- Étude des réseaux de reprise

Confrontation et exploitation des données(II)

Tirer (enfin !) tirer partie du corpus

- De meilleurs corpus pour mes étudiants déjà !
- Étude des noms de lieux
- Étude des réseaux de reprise
- Évaluation non supervisée de la qualité des données
- Attribution d'auteur
- Datation automatique

Références

References I

Barbaresi, A. (2021).
Trafilatura : A web scraping library and command-line tool for text discovery and extraction.
In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations*, pages 122–131.

Pincemin, B. (1999).
Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative.
In *Atelier Corpus et TAL : pour une réflexion méthodologique, Conférence TALN*, volume 99, pages 26–36.