Построение тематических моделей полилогов

Саттаров Тагир

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н К. В. Воронцов

Москва,
2021 г.

Цель работы

Задачи

- определить источники данных и привести их структурную характеристику.
- провести анализ метаданных запрошенных передач и сделать выводы об изменении в политике выпусков.
- выработать индикатор, основанный на анализе тональности, а также выявить границы его применения

Актуальность

- метод позволяет получить более обоснованные критерии изменений политики ведения телепередач
- анализ проводится в условиях отсутствия размеченных данных

Существующие работы

Схожие цели были поставлены в следующих работах

1. Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. *Language from police body camera footage shows racial disparities in officer respect(pnas.org/content/114/25/6521)*
2. Jean-Philippe Cointet, Sylvain Parasie« Ce que le big data fait à l'analyse sociologique des textes. Un panorama critique des recherches contemporaines », Revue française de sociologie 2018/3 (Vol. 59), p. 533-557. DOI 10.3917/rfs.593.0533

Новизна

На данный момент область исследования теледебатов остаётся слабо изученной:

1. не существует достаточного количества текстов в открытом доступе;
2. оцифровка передач ведётся с недавнего времени и существуют проблемы с точностью транскрипций.

Доступные выпуски

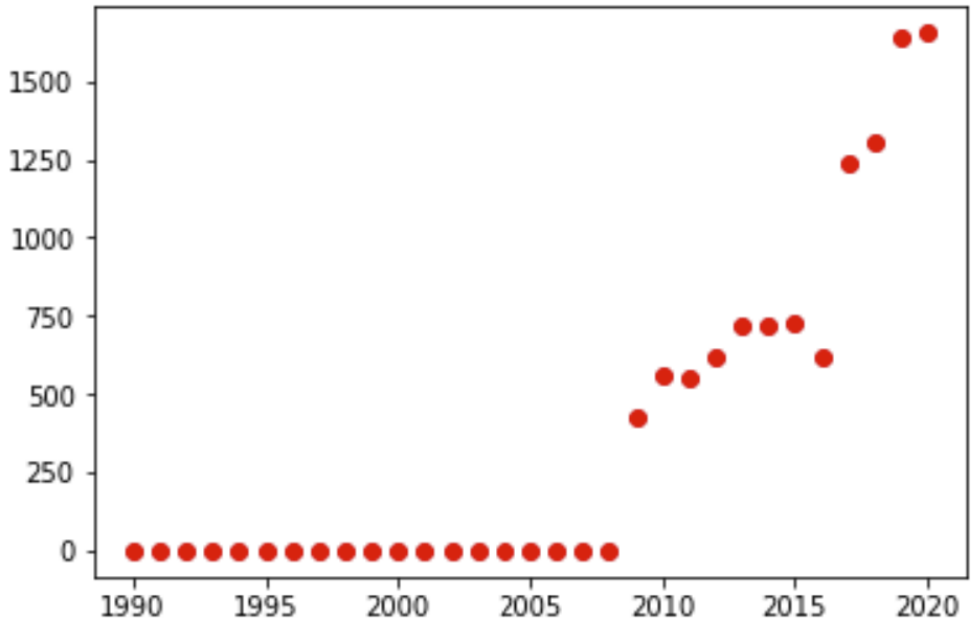


Рис.: Выпуски по годам

Выпуск	Кол-во	Канал	Суммарно
Journal de 20 heures	4368	TF1	4368
On n'est pas couché	433	France 2	433
C à vous C dans l'air	2378 3353	France 5	5731
Ça se dispute Face à l'info L'heure des pros On ne va pas se mentir	544 268 1339 551	Itélé/CNEWS	2702
Le Débat	1310	LCI	1310
Les Grandes Gueules	522	RMC	522

Рис.: Распределение выпусков по каналам

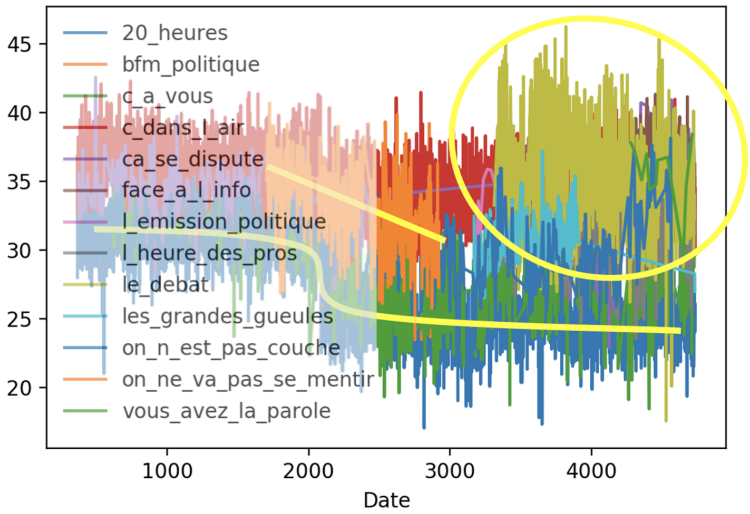
Структура данных и её особенности

<SpeechSegment ch="1"sconf="1.00"stime="830.820"
etime="832.670"spkid="S339"lang="fre"lconf="1.00"trs="1»
<Word id="2871"stime="830.82"dur="0.17"conf="0.70» que
</Word>
<Word id="2872"stime="831.09"dur="0.10"conf="0.41» et
</Word>
<Word id="2873"stime="831.20"dur="0.28"conf="0.41» eric
</Word>
<Word id="2874"stime="831.49"dur="0.35"conf="0.68»
menant </Word>
<Word id="2875"stime="831.87"dur="0.48"conf="0.81» raison
</Word>
</SpeechSegment>
Предложение – que et eric mennant raison

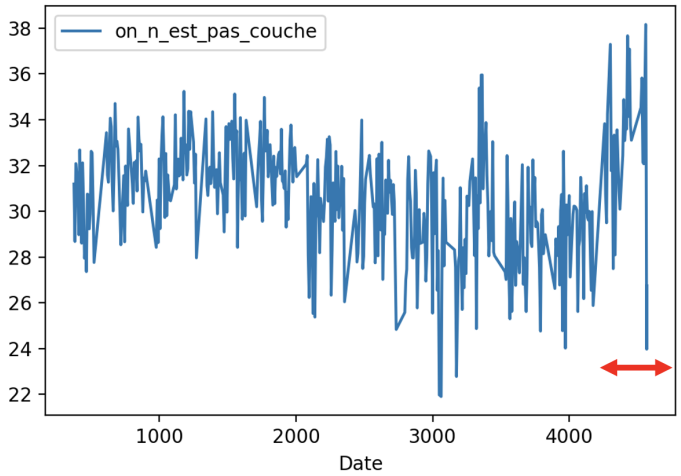
Анализ, основанный на длине фраз

Причины выбора

- неточность транскрипции слов не сказывается на результатах
- предполагается, что в более спокойных и размеренных полилогах средняя длина фразы больше.



(a) Средняя длина фразы для всех выпусков



(b) Средняя длина фразы для передачи on n'est pas couché

Анализ, основанный на длине фраз

Выводы

- изменение длины фраз в некоторых случаях может указывать на изменения в программе, такие как смена ведущего или политики канала
- недостатком является большой спектр факторов, от которых зависит средняя длина предложения

Анализ тональностей

Метод

Для анализа тональности используется метод TextBlob французской версии библиотеки TextBlob_fr. В качестве анализатора выбран PatternAnalyzer

Проблема

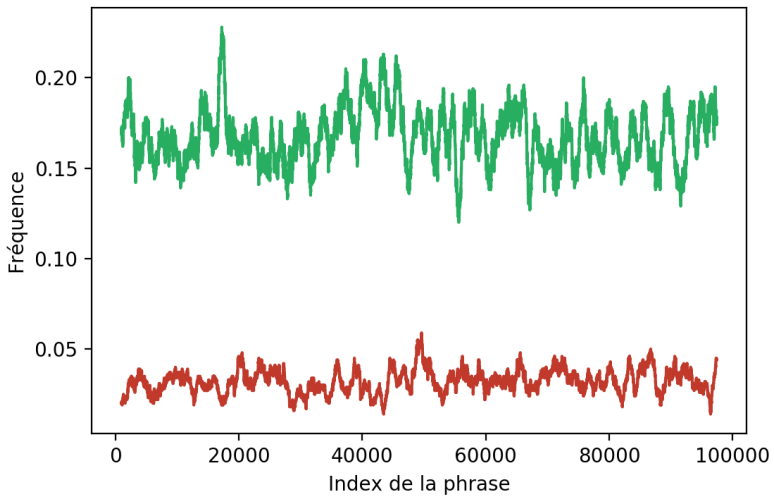
Главной проблемой является близость к нулю при подсчёте среднего значения тональности по передачам. Это может быть вызвано:

- неправильной транскрипцией слов;
- наличием рекламных или же новостных вставок;
- слишком большим количеством неизвестных слов.

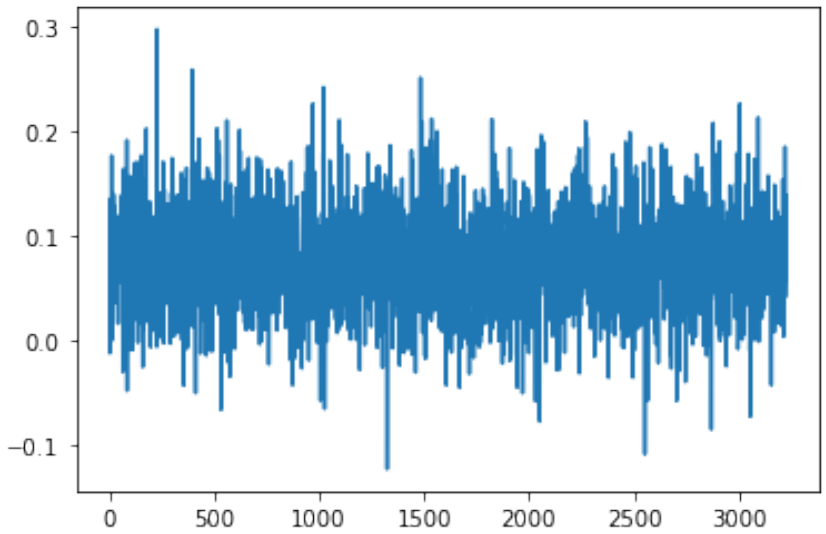
Анализ тональностей

Предложенные решения

- усреднение проводится для каждого "активного спикера"по отдельности
- учитываются отдельно негативные и позитивные тональности



(c) Частота негативных и позитивных тональностей для Face à l'info



(d) Тональности по участникам полилогов передачи L'Heure des pros

Анализ тональностей

Критика метода

- недостаточно полный набор правил и слишком длинные, не всегда точные транскрипции не позволяют модели точно оценивать тональность;
- спикеры, присутствующие в передачах, также могут быть определены неправильно.

Результаты, выносимые на защиту

- сделаны некоторые содержательные выводы о политике каналов на основе средней длины сегмента;
- выявлены недостатки метода оценки тональностей в задаче анализа транскрипций и проведены его усовершенствования.