Автоматический анализ тематики текста

от частотных слов — к вероятностному тематическому моделированию

Часть II: Тематическое моделирование

Даня Скоринкин, DH HSE

22 апреля 2020

Тематическое моделирование в Digital Humanities: кейсы

Кейс 1: литературный

Schöch C. Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama// Digital Humanities Quarterly. 2017. V. 011. № 2.

- Материал: корпус французской драмы с 1610 по 1810 гг. (проект *Théâtre classique*); на тот момент — 890 пьес
- Идея: исследовать связь между темами и жанрами
- Цитата: ..dramatic subgenres such as comedy or tragedy being in part defined on the basis of their themes, and Topic Modeling bringing to the fore the hidden thematic structure of text collections, it can be expected that Topic Modeling <…> should bring out relatively strong genre-related topic patterns in the data.

Отбор пьес в корпусе

- Не менее 3 и не боле 5 актов (явлений)
- Написана между 1630 и 1789 гг.
- Отнесена к одному из трех жанров:
  - Комедия
  - Трагедия
  - Трагикомедия

- **Итого**: 391 пьеса
  - 150 комедий, 189 трагедий, 52 трагикомедий
  - 5,6 млн токенов (слов)
  - 31 МБ в plain text

Предобработка

- Пьесы хранятся в TEI и могут быть разделены на явления
- Но автор решил делить просто на кусочки по 1000 слов
- Получилось 5840 фрагментов
  - (в среднем 14,9 на пьесу)
- Лемматизация и POS-теггинг: TreeTagger со специальной моделью от проекта PRESTO для анализа несовременного французского языка
- При помощи POS-теггинга отфильтровали:
  - Имена собственные
  - Все, что не является существительным, прилагательным или глаглом

Пайплайн эксперимента

Тематическое моделирование

- Метод: LDA
- Инструмент: MALLET (mallet.cs.umass.edu)
- Число топиков: 50, 60, 70, 80, 90, 100

Тематическое моделирование

- Метод: LDA
- Инструмент: MALLET (mallet.cs.umass.edu)
- Число топиков: 50, **60**, 70, 80, 90, 100

Тематическое моделирование

- Метод: LDA
- Инструмент: MALLET (mallet.cs.umass.edu)
- Число топиков: 50, **60**, 70, 80, 90, 100
  - Как выбрали это число?

Выбор числа топиков

«In order to evaluate the model and decide which model should be used <…>, a machine-learning task has been defined which consists of classifying the plays according to their (known) subgenre. This task appears to be appropriate since the main focus of the present study lies in the thematic differentiation of the dramatic subgenres <…>».

Результаты

Самые распространенные темы — обычно самые абстрактные или вообще ни о чем

«One common effect of a highly asymmetric topic probability distribution is that the topics with the highest probability scores (i.e. the ones most widely represented in the collection) are less interpretable than most others, in the sense that they are rather generic and vague».

Но в этой работе даже большие топики достаточно интерпретируемы:

«However, the model used here is notable for the absence of such topics; even the most widely present topics are relatively well-define»

Примеры топиков

“*coeur*, *amour*, *aimer*, *oser*, *madame*”
(heart, love, to love, to dare, madam)
Самый распространенный топик.
«Абстрактно-любовный»

“*coeur*, *amour*, *aimer*, *oser*, *madame*” (heart, love, to love, to dare, madam)
Самый распространенный топик. «Абстрактно-любовный»

“*secret*, *connaître*, *craindre*, *cacher*, *apprendre*” (secret, to know, to fear, to hide, to learn)
Тоже очень распространенный. «Топик интриги».

“*coeur*, *amour*, *aimer*, *oser*, *madame*” (heart, love, to love, to dare, madam) Самый распространенный топик. «Абстрактно-любовный»

“*secret*, *connaître*, *craindre*, *cacher*, *apprendre*” (secret, to know, to fear, to hide, to learn) Тоже очень распространенный. «Топик интриги».

“*vers*, *auteur*, *pièce*, *beau*, *esprit*” (verse, author, drama, beautiful, spirit/wit) «a metafictional topic centered on producing aesthetic literary works». В 10 раз менее распространен, чем абстрактно-любовный» (32).

“*coeur*, *amour*, *aimer*, *oser*, *madame*” (heart, love, to love, to dare, madam) Самый распространенный топик. «Абстрактно-любовный»

“*secret*, *connaître*, *craindre*, *cacher*, *apprendre*” (secret, to know, to fear, to hide, to learn) Тоже очень распространенный. «Топик интриги».

“*vers*, *auteur*, *pièce*, *beau*, *esprit*” (verse, author, drama, beautiful, spirit/wit) «a metafictional topic centered on producing aesthetic literary works». В 10 раз менее распространен, чем абстрактно-любовный» (32).

“*mal*, *monsieur*, *médecin*, *guérir*, *malade*” (hurting, mister, doctor, to recover/heal, sick). Самый редкий из 4-х. И самый конкретный: медицина.

Различия между топиками: «авторитарные» и «демократические» по составу

Доминируют
1-2 слова

Нет явно
доминирующих слов

Разнородность топиков:

- «What, then, are topics characteristic for this collection of plays, what are the themes most commonly found in them? <..> Many of the topics found are related to clear, abstract themes, such as love, death, crime and marriage»
- «Others focus on the dramatic personnel, such as topic 5 related to family members: “fils, père, mère, frère, enfant” (son, father, mother, daughter, child). Quite a number of topics are rather more concrete, such as those related to a setting, as in topic 46: “beau, berger, bois, arbre, lieu” (beautiful, shepherd, wood, tree, place). Others focus on quite specific activities, such as <…> topic 26 devoted to reading and writing letters: “lettre, lire, écrire, billet, main” (letter, to read, to write, note, hand)»

смерть

родство/персонажи

обстановка, природа, красоты

письмо

«..to put it another way, the “topics” found in literary texts are not only abstract themes such as justice, knowledge or crime, but also more concrete activities typically performed by fictional characters, like writing, eating, drinking, hunting, speaking and thinking, or literary settings or motives such as the sea, the forest, or interiors»

Связь тем и жанров

Как считали распределение тем по жанрам:

- «The topic scores obtained for each text segment are aggregated and their mean (or median) is calculated, taking into account the genre of the play that each text segment belongs to. This yields, for each topic, a score for its average importance, technically speaking its mean (or median) probability in each of the three subgenres»
- То есть для каждого фрагмента каждой пьесы известно содержание каждого топика – и приписан жанр пьесы

Topic Score Distribution
Most distinctive topics
amour-beau-yeux (53)
madame-aimer-monsieur (14)
bon-monsieur-beau (36)
sang-mort-main (39)
seigneur-soin-lieux (58)
monsieur-homme-chose (16)
coeur-amour-aimer (32)
coeur-ciel-malheureux (33)
peuple-loi-maître (56)
amour-coeur-aimer (28)
monsieur-marquis-bon (22)
madame-seigneur-coeur (54)
malheur-mal-sort (51)
trône-régner-empire (57)
secret-connaître-craindre (3)
monsieur-argent-bon (1)
combat-gloire-victoire (38)
amour-aimer-âme (37)
dessein-honneur-esprit (7)
homme-esprit-monde (13)
Comédie
Tragi-comédie
Tragédie
tc_subgenre
0.06
0.03
0.00
−0.03
−0.06

Topic Score Distribution
Most distinctive topics
amour-beau-yeux (53)
madame-aimer-monsieur (14)
bon-monsieur-beau (36)
sang-mort-main (39)
seigneur-soin-lieux (58)
monsieur-homme-chose (16)
coeur-amour-aimer (32)
coeur-ciel-malheureux (33)
peuple-loi-maître (56)
amour-coeur-aimer (28)
monsieur-marquis-bon (22)
madame-seigneur-coeur (54)
malheur-mal-sort (51)
trône-régner-empire (57)
secret-connaître-craindre (3)
monsieur-argent-bon (1)
combat-gloire-victoire (38)
amour-aimer-âme (37)
dessein-honneur-esprit (7)
homme-esprit-monde (13)
Comédie
Tragi-comédie
Tragédie
tc_subgenre
0.06
0.03
0.00
−0.03
−0.06

А есть ли динамика внутри пьес?

А есть ли динамика внутри пьес?

А если разбить пьесы на основе тем, как бы забыв о жанрах?

Эксперимент с data-driven кластеризацией французской драмы

«The differences observed so far rely on the *a priori* subgenre classification of plays, accepting the historical subgenre labels as given.

This, it may be argued, is problematic not only because the labels and their use may not be uncontroversial, but also because some seemingly significant differences in topic score distribution across such a small number of categories may always be found, no matter what the categories are».

Кластеризация методом главных компонент (PCA)

«When clustering plays based on their topic scores, then, tragicomedy appears to be more closely related to tragedy than to comedy. This fits one conventional description of tragicomedy as a tragedy that ends well, i.e. as a type of play with similar personnel, plot and themes as tragedy, except that there is no disastrous ending».

Иерархическая кластеризация

Нашлись интересные outliers (выбросы)

«The two tragedy clusters at the bottom of the graph do not contain a single play of a different subgenre label. But the occasional exception can be found in the three upper comedy clusters, where a small number of tragicomedies appear. One interesting case from among the more mixed clusters is Socrate, written by Voltaire in 1759 <…> and a rare specimen of a tragedy in prose, something which may explain its position separate from most other tragedies <…> In addition, although this play is labeled as a tragedy, it is certainly a very unusual tragedy and its historical subtitle is in fact “*comédie*”».

А угадываются ли жанры просто по частотным словам

- Да!
- И даже немного лучше, чем по топикам:

«It turns out that when comparing results from classification based on topic scores on the one hand and word frequencies on the other, performance for both methods is on a similar level, with a slight advantage for the word-based approach»

Выводы:

- Темы получились интерпретируемые
- Видимо, механизм выбора числа тем (а также гиперпараметра оптимизации) на основе жанра — сработал
- Жанровые особенности действительно выражаются через выделенные темы
- В некоторых случаях распределение темы по фрагментам пьес, видимо, отражает композиционные особенности
- Но для интерпретации всего этого очевидно нужен человек, понимающий, что в корпусе

Последователи:

EADH conference · Long Paper application

Topic Modeling 200 Years of Russian Drama

Irina Pavlova, Frank Fischer

(Higher School of Economics, Moscow)

Introduction

If the task was to explain Topic Modeling (TM) in one sentence, Wikipedia would have a good short answer: „In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract ‚topics‘ that occur in a collection of documents.“ (Wikipedia 2018) TM is a settled and popular method in the digital humanities and already reached a peak around 2012: „So many scholars in humanities departments are turning to [topic modeling] in their research that it is sometimes described as part of the digital humanities in itself“ (Schmidt 2012). And Matthew Jockers adds: „Topic models are the mother of all collocation tools.“ (Jockers 2013, p. 123)

Since its inception in the early 2000s, TM was meant to offer basic insights into large textual archives comprising tens of thousands of unstructured documents, promising to indicate topical patterns in otherwise unreadably big collections of texts. Since TM was developed with non-fiction document collections in mind, it took a while until it reached the literary studies. It took even longer until TM was applied on collection of dramatic texts. The study on French drama by Schöch (2017) and the forthcoming study on German drama by Reiter/Willand are rare exceptions. This paper sets out to establish a workflow for TM on Russian drama. The study is carried out using the Russian Drama Corpus (Skorinkin et al. 2018).

Topic #0
батюшка
матушка
отец
сударыня
жених
дочь
сын
жена
братец
ум
голова
господь
мать
часнилость
муж
государь
невеста
слава
брат
грех
воля
москва
правда

Topic #1
любовь
ум
ночь
жена
час
счастие
муж
женщина
тоска
нога
лицо
брат
москва
вздор
девушка
чувство
голова
мать
невеста
голос
страх
князь
беда
сила

Topic #2
сестра
дядя
отец
имение
деревня
старик
сын
ум
голова
москва
труд
лошадь
князь
брат
двор
слуга
закон
порядок
добро
грех

Topic #3
князь
царь
отец
народ
власть
сын
мир
смерть
честь
око
земля
дух
кровь
враг
государь
любовь
сила
небо
брат
гнев
слава
страсть
надежда

Topic #4
деньги
тысяча
барин
старик
рубль
земля
баба
правда
город
мужик
жена
грех
братец
отец
совесть
собака
дорога
честь
голова

Topic #5
господин
минута
случай
ночь
место
город
сад
сторона
женщина
муж
письмо
лицо
комната
вечер
удовольствие
положение

Topic Distribution by Author

We chose plays by 15 authors of our corpus to visualise topic distribution in their works. The probability values show the mean of all values per topic per play by the same author (fig. 4).