Софийски университет „Св. Кл. Охридски”

Факултет по математика и информатика

Курсов Проект

на тема:

Клъстеризация на новинарски текстове

Студент: Мелания Стоянова Бербатова Ф.Н. **26032**

Курс: „Откриване на знания в данни“,Учебна година: 2018/19

Преподаватели: **проф. Иван Койчев**

Декларация за липса плагиатство:

- Плагиатство е да използваш, идеи, мнение или работа на друг, като претендираш, че са твои. Това е форма на преписване.
- Тази курсова работа е моя, като всички изречения, илюстрации и програми от други хора са изрично цитирани.
- Тази курсова работа или нейна версия не са представени в друг университет или друга учебна институция.
- Разбирам, че ако се установи плагиатство в работата ми ще получа оценка “Слаб”.

25.6.19 г. Подпис на студента:

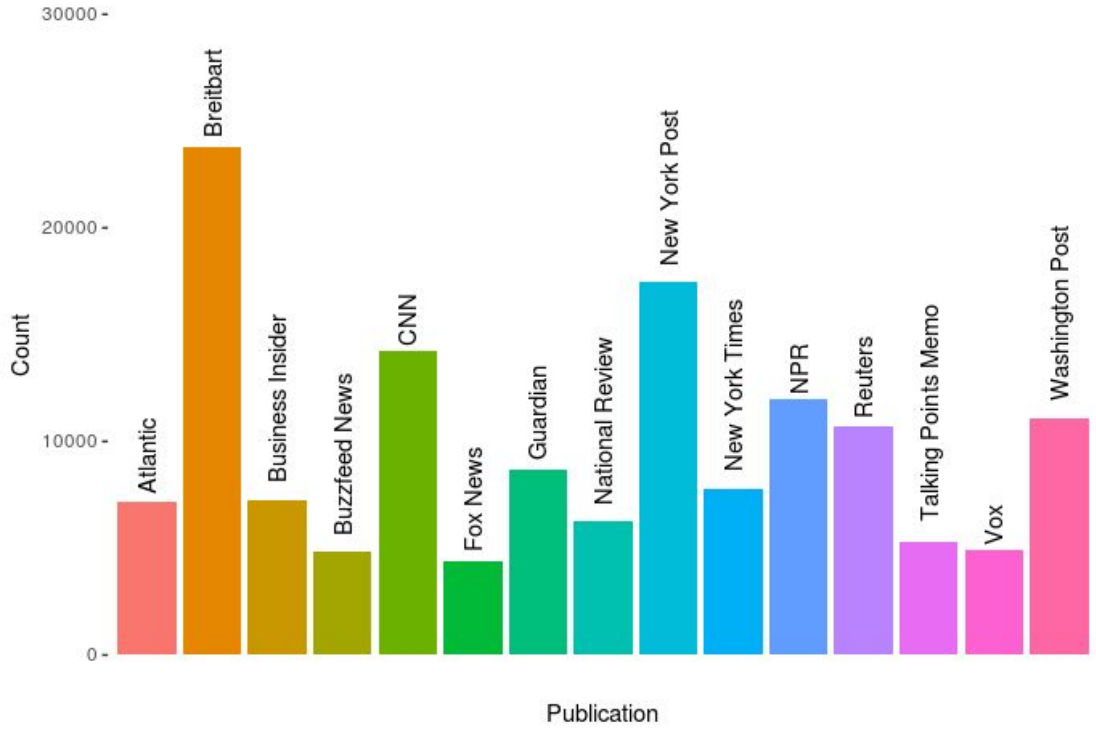
1. Описание на данните

За целите на проекта са използвани данните „All the news”, налични свободно на платформата Kaggle.com. Авторът на публикацията е събрал новинарски данни от New York Times, Breitbart, CNN, Business Insider, the Atlantic, Fox News, Talking Points Memo, Buzzfeed News, National Review, New York Post, the Guardian, NPR, Reuters, Vox, and the Washington Post. Новините са избирани така, че да бъдат с предимно политическа насоченост, и да показват различни политически възгледи.

Повечето от новините датират от 2016 до юли 2017, макар че има и новини и от преди това.

Новините са събирани чрез библиотеката *BeautifulSoup* и са налични в три *.csv* файла. Общия обем е около 150 000 новини. Поради големия им обем, за някои алгоритми ще използваме само част от тези данни.

Разпределението на новините по източници може да бъде видяно на следната графика:



Фиг. 1. - Разпределение на новините по източници

Данни

На следващата фигура са показани данните, прочетени чрез Dataframe

	Unnamed: 0	id	title	publication	author	date	year	month	url	content
0	0	17283	House Republicans Fret About Winning Their Hea...	New York Times	Carl Hulse	2016-12-31	2016.0	12.0	NaN	WASHINGTON — Congressional Republicans have...
1	1	17284	Rift Between Officers and Residents as Killing...	New York Times	Benjamin Mueller and Al Baker	2017-06-19	2017.0	6.0	NaN	After the bullet shells get counted, the blood...
2	2	17285	Tyrus Wong, 'Bambi' Artist Thwarted by Racial ...	New York Times	Margalit Fox	2017-01-06	2017.0	1.0	NaN	When Walt Disney's "Bambi" opened in 1942, cri...
3	3	17286	Among Deaths in 2016, a Heavy Toll in Pop Musi...	New York Times	William McDonald	2017-04-10	2017.0	4.0	NaN	Death may be the great equalizer, but it isn't...
4	4	17287	Kim Jong-un Says North Korea Is Preparing to T...	New York Times	Choe Sang-Hun	2017-01-02	2017.0	1.0	NaN	SEOUL, South Korea — North Korea's leader, ...

Табл.1. Поглед върху данните

Описателни статистики

На табл. 1 са показани описателни статистики относно броя символи, използвани в новините и в техните заглавия. Тези статистики ни дават обща информация за разнообразието на данните, а така също и за насока на избора на подходяща дължина за автоматично генерирани резюмета.

Statistics	**Length of summary**	**Length of text**
mean	62.02	4700.51
std	17.89	4526.59
min	1	1
25.00%	51	2296
50.00%	62	3883
75.00%	74	5851
max	171	292586

Табл. 1. Описателни статистики на броя символи, използвани в наличните данни

2. Клъстеризация на новини

Целта на настоящата секция е да определим основните групи новини, клъстери, които са налични в използвания набор от данни, както и да оценим поведението на различни клъстеризиращи алгоритми върху данните.

Важно уточнение е, намирането на оптимални клъстери може и да не съвпада с намирането на най-добре интерпретируемите теми. Доколко най-добре дефинираните теми съвпадат с най-добре оценените клъстеризации е предмет на проучване на настоящия проект.

Подготовка на данните

Първо, данните се превръщат във tf-idf вектори.

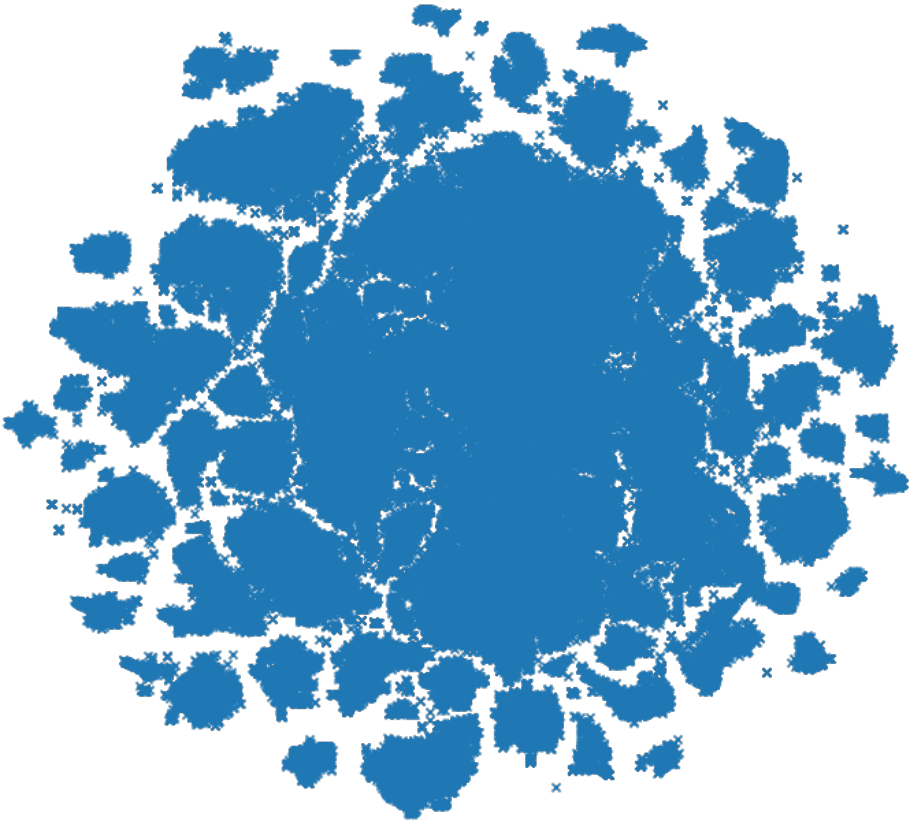
Първоначално намаляне на размерността чрез LSA

Tf-idf векторите обаче са разпръснати вектори с много размерности, поради което са непосилни за визуализация. За това първо се предават на алгоритъм за латентен семантичен анализ (LSA), който да намали размерността на 50.

Намаляване на размерността и визуализиране

С цел визуализация на обектите, най-добре е те да бъдат превърнати в обекти в размерност 2. Класически алгоритъм за това е Principal component analysis (PCA). PCA обаче не работи добре за данни, които имат нелинейни зависимости, каквито са характерни за данните от текстови формат. За данни с нелинейни зависимости са разработени много алгоритми, един от които е t-SNE (t-distributed stochastic neighbor embedding). Той създава вграждания на обектите в по-малко пространство, като запазва локалните близости.

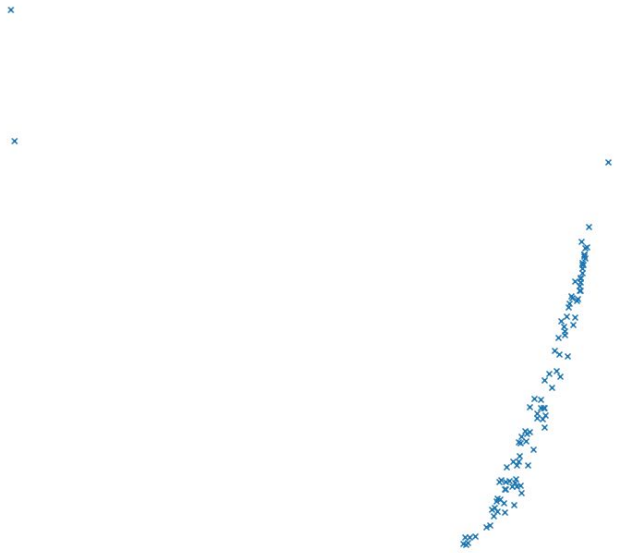
На следващата графика са показани данните, редуцирани до размерност 2 чрез алгоритъма t-NSE със 121 съседа. Можем да забележим един доминиращ клъстер в средата (както по-късно ще разберем, съставен от новини, свързани с президентските избори в САЩ), както и голямо множество по-малки клъстери.



Фиг. 2. Визуализация на данните чрез алгоритъма t-NSE.

Визуализация чрез метода за да Локално линейно вграждане (LLE)

На следващата фигура са визуализирани данните чрез друг метод за намаляване на размерността при нелинейни зависимости - метода за локално линейно вграждане. За съжаление, въпреки че метода е в пъти по-бърз, не помага за доброто визуализиране на данните:



Фиг. 3. Визуализация на данните чрез алгоритъма LLE.

Базов метод – k-Means

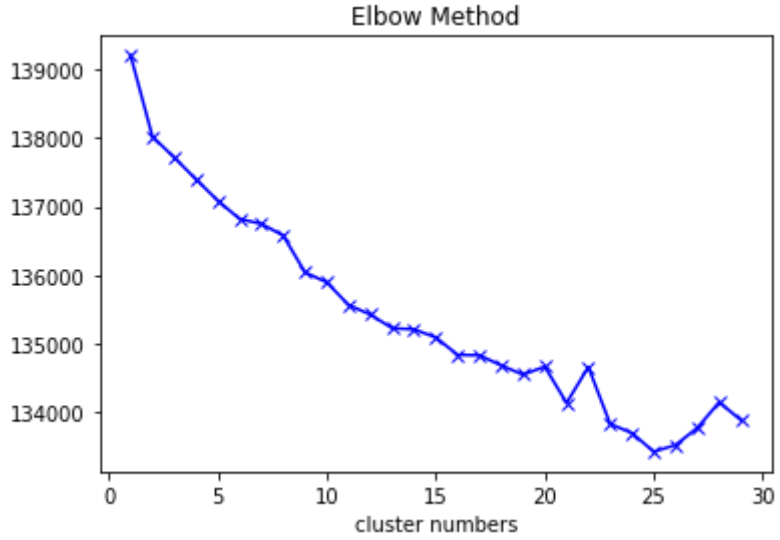
K-Means е най-популярния метод за клъстеризация. За да го използваме ефективно обаче е необходимо предварително да уточним стойността на параметъра k- необходимия брой клъстери.

Това може да стане посредством метода Elbow. Той гласи следното:

„Броят на клъстери трябва да бъде избран така, че добавянето на друг клъстер да не дава много по-добро моделиране на данните.“

В термините на k-Means това означава, че ако при добавянето на клъстер, средноквадратичната грешка на данните не намалява, или намалява малко, сме избрали оптимален брой клъстери.

За да приложим на практика метода, извикваме k-Means с различен брой клъстери и наблюдаваме движението на средноквадратичната грешка:



Фиг. 4. Визуализация на методът Elbow за избиране на стойността на параметъра k

Виждаме, че най-рязкото падане в стойностите на средноквадратичната грешка е до към k=4, като има още едно сравнително стръмно спускане до към k=8. След k=15 стойностите нямат постоянен спад. На базата на тези наблюдения, можем да направим клъстеризиране с k=4 и k=8 и да сравним резултатите.

При избор **к=4** получаваме следните клъстери. Със скоби в курсив са примерни наименования, които могат да бъдат зададени на клъстерите.

cluster0: (*US president elections*)
['trump', 'clinton', 'said', 'president', 'campaign', 'republican', 'donald', 'hillary', 'house', 'obama', 'mr', 'sanders', 'election', 'party', 'cruz']

cluster1: *(not meaningful)*
['says', 'like', 'people', 'just', 'women', 'said', 'time', 'don', 'think', 'new', 'know', 'life', 'way', 'game', 'really']

cluster2: (*US international politics*)
['said', 'syria', 'trump', 'korea', 'military', 'united', 'russia', 'president', 'islamic', 'north', 'syrian', 'isis', 'obama', 'state', 'iran']

cluster3: *(US national politics)*
['said', 'police', 'new', 'people', 'percent', 'court', 'company', 'trump', 'state', 'news', 'year', 'president', 'according', 'law', 'told']

При **к=8** получаваме следните резултати:

cluster0:

['people', 'like', 'says', 'women', 'just', 'said', 'don', 'time', 'think', 'new', 'know', 'life', 'really', 'way', 've']

cluster1:

['court', 'supreme', 'justice', 'trump', 'law', 'order', 'judge', 'said', 'gorsuch', 'federal', 'president', 'immigration', 'scalia', 'ban', 'obama']

cluster2:

['said', 'police', 'new', 'people', 'percent', 'company', 'year', 'according', 'news', 'told', 'state', 'city', 'million', 'years', 'time']

cluster3: **education**

['students', 'school', 'schools', 'education', 'student', 'university', 'campus', 'college', 'said', 'teachers', 'devos', 'colleges', 'kids', 'public', 'parents']

cluster4:

['clinton', 'sanders', 'hillary', 'trump', 'campaign', 'democratic', 'said', 'voters', 'state', 'presidential', 'emails', 'fbi', 'comey', 'obama', 'bernie']

cluster5:

['trump', 'republicans', 'obamacare', 'republican', 'house', 'cruz', 'health', 'democrats', 'senate', 'said', 'tax', 'rubio', 'care', 'party', 'gop']

cluster6:

['trump', 'said', 'donald', 'president', 'campaign', 'republican', 'mr', 'clinton', 'cruz', 'people', 'obama', 'presidential', 'election', 'house', 'white']

cluster7:

['trump', 'russia', 'said', 'russian', 'president', 'obama', 'korea', 'syria', 'intelligence', 'north', 'iran', 'military', 'united', 'nuclear', 'syrian']

Забелязваме, че се получават и нови тематики, като тази за образованието.

При увеличение на броя центроиди на 15 темите стават още по-ясно разграничими, като например разделението на темата за международната политика на тези 3 клъстера:

cluster2:

['syria', 'isis', 'said', 'syrian', 'iran', 'islamic', 'military', 'forces', 'turkey', 'state', 'iraq', 'russia', 'assad', 'war', 'attack']

cluster11:

['korea', 'north', 'korean', 'missile', 'nuclear', 'kim', 'china', 'south', 'pyongyang', 'said', 'jong', 'trump', 'missiles', 'test', 'ballistic']

cluster14:

['comey', 'trump', 'russia', 'intelligence', 'russian', 'fbi', 'flynn', 'investigation', 'president', 'said', 'house', 'committee', 'putin', 'election', 'director']

но получаваме и по-голям брой безсмислени клъстери като:

cluster13:

['said', 'new', 'people', 'company', 'court', 'like', 'year', 'time', 'news', 'just', 'years', 'state', 'told', 'president', 'says']

Оценяване на к-Means - коефициент на силуета.

Резултатите от k-means можем да оценим както ръчно (“на око”), така и чрез метрики, оценяващи клъстеризацията. Коефициентът на силуета се изчислява, като се използва средното междуклъстерно разстояние (a) и средното най-близко разстояние до другите клъстери (b) за всеки клъстер. Коефициентът на силуета за клъстер е *(b - a) / max (a, b)*.

Коефициентът на силуета дава оценка доколко отделните клъстери са достатъчно добре обособени. Най-добрата стойност е 1, а най-лошата стойност е -1. Стойностите в близост до 0 показват припокриващи се клъстери.

При k=4 получаваме оценка 0.0014. При k=8 оценката е 0.0012 И при k=15 е -0.0041, което означава, че при повече клъстери границите се размиват.

DB-scan

Въпреки, че DBScan не се препоръчва като алгоритъм за клъстеризация на текст, за пълнота е включен в настоящия проект. Коефициентът на силуета на DB-scan е -0.28, което е значително по-ниска стойност от всички стойности за k-means.

Откриване на теми в текст

Моделирането на теми (Topic modelling) в текстови документи е задача, много близка до задачата за намиране на клъстери в текстови документи. Най-разпространения алгоритъм за това е LDA - Латентен анализ на Дирихле. Заради големия обем на данните обаче за нашата задача е по-подходящо да се използва друг алгоритъм - NMF.

Резултатите от проведения експеримент за моделиране на 4 тема са много близки до тези от алгоритъма k-means с 4 центроида.

Topic0:
['presidential', 'obama', 'party', 'gop', 'election', 'white', 'said', 'republicans', 'mr', 'house', 'cruz', 'campaign', 'republican', 'donald', 'president']

Topic1:
['going', 'company', 'think', 'don', 'health', 'years', 'time', 'year', 'women', 'new', 'just', 'says', 'like', 'percent', 'said']

Topic2:
['candidate', 'email', 'percent', 'election', 'party', 'fbi', 'presidential', 'emails', 'bernie', 'state', 'voters', 'campaign', 'democratic', 'hillary', 'sanders']

Topic3:
['korea', 'government', 'united', 'russian', 'officials', 'north', 'military', 'security', 'obama', 'president', 'syria', 'state', 'russia', 'mr', 'police']

3. Заключение

За намаляване на размерността и визуализация на многоразмерни данни с нелинейни зависимости, каквито са новините, е най-добре да се използва комбинация от латентен семантичен анализ и t-NSE. Най-добре обособени клъстери се получават от алгоритъма k-means при малко k, а при по-голямо k групирането по теми е по-добре обособено.