МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева (Самарский университет)»

Институт информатики, математики и электроники

Факультет информатики   
Кафедра программных систем

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

к лабораторной работе №3 по дисциплине «Большие данные»

по теме «Анализ содержимого большого количества документов

методами кластеризации»

Студент группы 6132-020402D В.А. Артамонов

Руководитель В.В. Жидченко

Самара 2021

СОДЕРЖАНИЕ

[1. Словесное описание 5](#_Toc91853037)

[2. Зависимость качества разделения кластеров от используемых параметров алгоритмов 6](#_Toc91853038)

[3. Результаты кластеризации 9](#_Toc91853039)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 13](#_Toc91853040)

Описание и анализ предметной области

В данной лабораторной работе необходимо кластеризовать корпус документов, состоящий из англоязычных статей, содержащих словосочетание «big data». Это позволит разбить множество статей на отдельные темы, использующие указанное словосочетание. В результате можно получить представление о различных направлениях исследований в области Big Data.

Анализируемый корпус документов состоит из 2141 статьи, которые были опубликованы в период с 2016 по 2020 г. и получены в результате поискового запроса «big data» на сайте модерируемого архива научных статей с открытым доступом arXiv.org (ARXIV, 2018).

Задание:

1. Скачать архив с содержимым статей. Архив содержит статьи в формате TXT, полученные автоматическим преобразованием из формата PDF для упрощения анализа. Имена файлов имеют формат YYMM.N.txt, где YY – год публикации статьи, MM – месяц публикации, N – порядковый номер статьи в пределах соответствующего месяца.
2. Выполнить кластеризацию корпуса документов, используя библиотеки scikit-learn и nltk языка Python. Для этого:

2.1) Создать dataset из файлов скачанного архива в формате TXT;

2.2) Используя библиотеку nltk, выполнить стемминг документов, образующих dataset;

2.3) Выполнить кластеризацию корпуса документов с использованием словаря стоп-слов английского языка и алгоритма K-means. При необходимости дополнить словарь стоп-слов самыми часто используемыми словами из области Big Data. Сначала необходимо выполнить кластеризацию несколько раз, задав число кластеров равным 12, и изменяя различные параметры алгоритма K-means. Выбрать те параметры, которые обеспечили наилучшее разделение кластеров, используя один из критериев качества разделения, например, анализ силуэта. Затем необходимо отдельно выполнить кластеризацию того набора документов, который сформировал самый объемный кластер, т. е. тот из 12-ти найденных кластеров, который содержит максимальное количество документов. При этом количество искомых кластеров задать равным 6.

2.4) Определить тему статей, принадлежащих найденным кластерам. Для этого воспользоваться списком наиболее часто встречающихся слов в каждом кластере, а также просмотреть названия и бегло просмотреть содержание нескольких статей из каждого кластера, пользуясь файлами со статьями. При необходимости можно скачать любую статью с сайта arxiv.org в формате PDF.

1. Составить отчет по результатам работы.
2. Словесное описание

Программный алгоритм можно разделить на следующие части:

1. Загрузка стоп-слов – массива слов, которые не будут учитываться в стемминге и кластеризации.
2. Выбор стеммера (PorterStemmer, SnowballStemmer). Второй алгоритм во многом превосходит первый, в то время, как первый – только на английском.
3. Инициализация командной строки.
4. Загрузка статей и выделение основ слов для их дальнейшей кластеризации.
5. Расширение сформированного списка стоп-слов, в который добавляются специфичные для научных работ слова.
6. Выделение признаков (features) из исходных данных. Для этого используется один из двух векторизаторов. Первый (HashingVectorizer) конвертирует коллекцию исходных документов в матрицу из токенов обнаружения слов. Второй (TfidfVectorizer) конвертирует коллекцию исходных документов в матрицу TF-IDF признаков.
7. При включении алгоритма LSA - уменьшение размерности задачи.
8. Выбор алгоритма кластеризации (MiniBatchKMeans или KMeans).
9. Расчёт центроидов.
10. Вывод информации качества кластеризации.
11. Определение кластера для каждого входного документа.
12. Формирование для каждого кластера файла, в который записывается список статей, входящих в него.
13. Вывод списка наиболее часто встречающихся слов.
14. Зависимость качества разделения кластеров от используемых параметров алгоритмов

Коэффициент силуэта – используется для анализа качества кластеризации [1].

Рисунок 1 – Зависимость качества кластеризации по «--no-minibatch»

Рисунок 2 – Зависимость качества кластеризации по «--no-idf»

Основным преимуществом Minibatch K-means, жертвуя малой точностью разбиения, происходит уменьшение вычислительных затрат на поиск кластеров [2].

При использовании коэффициента силуэта, (при включении IDF) можно заметить ухудшение качества кластеризации, так как, присущее коэффициенту «проклятие размерности» для наборов данных высоких размеров (текстовые наборы данных).

* «--use-hashing» - параметр, который уменьшает объем используемой памяти и несколько ускоряет обработку данных, но на качество кластеризации не оказывает никакого влияния.
* «--stem» - параметр, который не варьировался в рамках работы, так как его отключение приводило к невозможности адекватной кластеризации набора документов.
* «--clusters» - параметр, который не варьировался, так как был задан.
* «--lsa» - параметр, который включает латентно-семантический анализ и увеличивает скорость работы программы [3].

Рисунок 3 – Зависимость качества кластеризации по «--n-features»

С увеличением параметра «--n-features» качество кластеризации ухудшается. Это связано с тем, что при увеличении размерности задачи увеличивается сложность кластеризации набора данных.

В результате проведенных экспериментов были выбраны следующие параметры: --no-idf, --no-minibatch, --n-features 10000, --clusters 12 (6) [4].

1. Результаты кластеризации

Все статьи были разбиты на 12 кластеров.

Рисунок 4 – Структура кластеров

Перечень самых часто употребляемых слов из этих кластеров представлен в таблице 1.

Таблица 1 – Список самых часто употребляемых слов в 11 основных кластерах

|  |  |
| --- | --- |
| Название кластера | Часто употребляемые слова |
| 1 | 2 |
| Privacy | detection awareness intrusion trusted privacy data ﬁltering security prevention ownership social cyber IoT aggregation analytics service blockchain |
| Complex networks | networks path complex wireless software compact multilayer structure IoT graph architecture node visualisation sensor analysis optimization allocation simulation |
| Graph processing | graph distributed partitioning compression analysis visualization dynamic network neighborhood extraction aggregation pattern relation learning classiﬁcation database parallel allocation |
| Object recognition | object recognition 3D GAN archives image parallel context tracking deepfakes vision semantic modeling 2D retrieval reconstruction quality scalable point |

Продолжение таблицы 1

|  |  |
| --- | --- |
| 1 | 2 |
| Clustering | clustering measurements data dimensional mining centroid empty k-means graph frequency unstructured big curve pattern diffusion learning vector trajectories mapreduce texts |
| Mathematical methods | algorithm stochastic optimization randomized matrix parallel multinomial regression bounded gradient approximation estimation error optimal quantum order rank positive maximal allocation |
| Data processing algorithms | analysis methods techniques data mining hashing batch tuple processing frequent classiﬁcation projection classes metadata stream normalization redundancy complexity extraction search |
| Healthcare | health social statistics bio medicine privacy recommender effect literature precision clinical decision ethics dynamic urban modelling research studies world analytics |
| Natural science research | signal learning measurement electro quantum reconstruction ﬁelds classiﬁer science statistical ethics celestial mechanics analysis data stars sun evolution chemically activity |
| Statistics | statistics bayesian gaussian empirical process hierarchical model nonlinear correlation dimension dependence average functional random asymptotic distribution inverse regression series kurtosis |
| Distributed computing | distributed computing big data high networks apache hadoop performance cloud scheduler throughput modelling benchmark benchmark parallel analytics mining grid cluster |

В самом большом из 12 кластеров находилось 389 статей. После его дальнейшего разбиения получился следующий набор кластеров (рисунок 5).

Рисунок 5 – Структура самого большого кластера

Список самых часто употребляемых слов в документах кластера представлен в таблице 2.

Таблица 2 – Список самых часто употребляемых слов

|  |  |
| --- | --- |
| Название кластера | Часто употребляемые слова |
| General Big Data Research | use th set valu time comput number method process function perform model inform algorithm result approach pp base point |
| Social Research | social time network studi use dataset th number user citi model differ mobil analysi result inform visual predict work urban |
| Articles in Italian | los ny etal newyork para por las ao dato tweet cë com moral como word es twitter acm usa caso |
| Networking | pp network acm smart ou grid howev doi theâ scale permission tools http org ieee net graph www cloud |
| Recommendation Systems | user recommend time set algorithm network use item twitter th model inform result similar number method rate social differ al |
| Astronomy | al scale studi diï photon number use th time valu analysi high measur method result state observ eld model light |

Рисунок 6 – Доля документов, образующих каждый кластер

Рисунок 7 – Зависимость количества документов от года публикации

Зависимость доли документов, образующих каждый кластер, по отношению к общему числу документов в корпусе, от года публикации представлена на рисунке 8.

Рисунок 8 – Зависимость доли документов, образующих каждый кластер, по отношению к общему числу документов в корпусе, от года публикации

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. 2.3 Clustering – scikit – learn 1.0.2 documentation [Электронный ресурс] URL: <https://scikit-learn.org/stable/modules/clustering.html#k-means> (дата обращения 31.12.2021);
2. sklearn.cluster.KMeans – scikit-learn 1.0.2 documentation [Электронный ресурс] URL: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMe> (дата обращения 31.12.2021);
3. 6.2. Feature extraction – scikit-learn 1.0.2 documentation [Электронный ресурс] URL: <https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction> (дата обращения 31.12.2021);
4. Анализ содержимого большого количества документов методами кластеризации: Метод. указания к лабораторным занятиям / Самар. ун-т; Сост. В.В. Жидченко. Самара, 2020. 10 с.