Формирование визуализированного представления патентного ландшафта

Д.М. Коробкин, М.В. Савельев, С.А. Фоменков, Г.А. Верещак

Волгоградский государственный технический университет, Волгоград

Аннотация: Рассмотрены и использованы методы и технологии для решения задачи визуализации патентного ландшафта на основе кластерного анализа патентного массива. Разработаны алгоритмы загрузки патентных архивов, парсинга патентных документов, кластеризации патентов и визуализации патентного ландшафта. Реализован программный модуль для кластеризации патентных документов на основе модели латентного размещения Дирихле и визуализации патентного ландшафта на данных кластеризации с использованием библиотек gensim, PySpark, sklearn. Программный модуль апробирован на патентах, выданных ведомством по патентам и товарным знакам США.
Ключевые слова: патенты, извлечение информации, кластеризация, патентный ландшафт, инновационный потенциал.

Введение

Процесс генерации новых технических решений, поиск аналогов на разработанную технологию и анализ технологических трендов могут быть упрощены за счет анализа существующей патентной базы [1, 2]. В патентном документе закреплены ключевые особенности изобретения, которые делают его уникальным, а также описываются процессы, необходимые для его воспроизведения. Согласно статистическим данным с каждым годом количество запатентованных технологий только увеличивается, поэтому найти среди множества патентных документов те, за которыми закреплена технология с высоким инновационным потенциалом и перспективностью становится все сложнее.

Говоря о задаче определения перспективности и инновационного потенциала технологии, закрепленной в патенте, можно утверждать, что какого-либо эталона решения и автоматизации подобной задачи нет.

Достаточно распространенным способом анализа технологии, представленной в патенте, является метод SAO. В [3] авторы проверяют близость технологии к технологическому тренду посредством извлечения

матричных векторов, состоящих из структур SAO. А в [4] создаются терм-документные матрицы на основе структур SAO, где документами являются патенты, а термами - структуры SAO. В таком случае инновационный потенциал представлен относительным размером тематического кластера. Так же модель SAO для извлечения информации из англоязычных патентов используется в [5-7]. В [8] отмечается необходимость изучения структуры патентного документа для повышения качества извлечения данных.

В [9] авторами предложен "Инновационный радар". Это метод определения инновационного потенциала в виде опросного листа по патенту. Заполнив опросный лист, авторы получают оценку потенциала патента по 100-балльной шкале.

В [10] авторы анализировали пользовательские отзывы к патенту для определения инновационного потенциала. Из пользовательских отзывов можно извлечь информацию о технологии, указанной в патенте, и окрас реакции пользователей.

Авторы [11] создали методику для определения технологий с высоким потенциалом - Technology Opportunity Discovery. В этой работе используется кластеризация патентов по ключевым словам. Согласно [11], авторы способны выделить технологию с высоким потенциалом из множества.

Цель работы - разработка программного модуля для кластеризации патентных документов USPTO на основе модели латентного размещения Дирихле (LDA) и визуализации патентного ландшафта для повышения эффективности выявления новых тенденций и идей в области синтеза технических решений.

Материалы и методы

Для выполнения задачи кластеризации патентных документов и визуализации патентного ландшафта необходимо выполнить ряд

подготовительных действий. К таким действиям относятся загрузка и анализ патентных массивов.

Входными данными алгоритма загрузки патентных архивов являются ссылка на веб-страницу с патентными архивами ведомства по патентам и товарным знакам США (USPTO) за выбранный год и абсолютный путь к директории, в которую необходимо загрузить и разархивировать патентные архивы. Веб-страница представляет собой перечень ссылок на патентные архивы, которые публикуются еженедельно. Патентный архив представляет собой ZIP-файл, содержащий патентный массив в XML-файле. По веб-странице выполняется поиск определенных тегов HTML, содержащих ссылки на скачивание патентных архивов. Каждая найденная ссылка добавляется в список ссылок для дальнейшей обработки. Для ссылок в сформированном списке выполняются операции скачивания архива, его разархивирования и сохранения в указанную директорию. Выходными данными алгоритма являются разархивированные недельные патентные массивы в XML-файлах.

На вход алгоритма парсинга патентных документов поступает путь к директории, содержащей недельные патентные массивы, полученные в ходе работы алгоритма загрузки патентных архивов. Патентный массив представляет собой XML-файл, внутри которого хранятся описания патентов в формате XML, где перед каждым патентом происходит объявление того, что это - XML. Такое объявление допустимо только в начале XML-файла, поэтому первоначальный XML-файл патентного массива не является синтаксически правильным и существует необходимость в разделении патентного массива на единичные патентные документы.

Указанная директория анализируется на наличие вложенных папок с патентными массивами. Найденные патентные массивы разбиваются на

отдельные патентные документы в виде XML-файлов. Для каждого патентного документа осуществляется парсинг.

Патент проверяется на наличие тега *classification-ipcr*, содержащего описание класса патента согласно Международной патентной классификации (МПК). Если описание класса найдено, осуществляется дальнейшая обработка документа, производится парсинг таких полей патента, как: *section*, *class*, *subclass*, *main-group*, *description*. Для соответствия патента заданной категории проверяются поля *section* и *class*, и если проверка успешна, то извлекаются данные тегов *subclass*, *main-group* и *description*. Множество вложенных тегов p элемента *description* является описанием патента. Текстовые узлы вложенных тегов *p* объединяются для формирования единого текста с описанием патента. Извлеченная информация записывается в csv-файл, являющийся коллекцией отфильтрованных патентов. Данный анализ проводится с каждым патентным документом.

Алгоритм формирования набора данных патентных документов предназначен для формирования набора данных патентов, отфильтрованным по необходимым категориям для последующей кластеризации, и их количества по срезу каждой категории. Алгоритм считывает объемные входные данные, размер которых может превышать несколько гигабайт, а затем преобразовывает их в формат *parquet*.

Входными данными является коллекция патентов в формате *csv*, полученная в результате алгоритма парсинга патентных документов. Выходными данными является файл формата *parquet* со следующими столбцами: группа данного патента ("*main-group*" в классификации МПК), подкласс данного патента ("*subclass*" в классификации МПК), текст описания патента, количество слов в описании патента.

Для работы алгоритма формирования набора данных патентных обязательно наличие столбца с текстом описания патента. Перед преобразованием в *parquet* необходим ряд действий: строки входных данных проверяются на нулевые значения, над столбцами выполняются функции, определенные пользователем (*user-defined functions*), патенты сортируются по убыванию числа слов в тексте описания патента. Полученный файл может быть отфильтрован по принадлежности к категории классификации и/или числу патентов в выбранной категории, после чего будет создан новый файл формата *parquet*. Также алгоритм предусматривает объединение нескольких выходных файлов *parquet* для получения композитного набора данных.

Для построения патентного ландшафта необходимо выполнить нормализацию текстов с описанием патентов, кластеризовать наборы данных патентов методом LDA, визуализировать результаты работы модели LDA.

Нормализация текстов описаний включает в себя токенизацию текста, лемматизацию токенов, удаление стоп-слов и векторизацию токенов, составляющих текст описания патента. Результатом нормализации являются словарь токенов и векторизованное представление набора патентов. Нормализацию необходимо выполнять, чтобы модель LDA смогла корректно кластеризовать данные.

Результаты работы модели машинного обучения визуализируются, составляя из совокупности графиков патентный ландшафт. Входными данными для модели LDA является набор данных со столбцом векторизованных документов и словарь с токенами. Патентный ландшафт включает в себя таблицу распределения категорий патентов по кластерам, графическое представление кластеров на координатной плоскости, график выявленный ключевых слов в каждом кластере и их вес, график распределения количества токенов в каждом документе, график распределения количества токенов в каждом документе по срезу кластера,

графическое представление кластеров на координатной плоскости с помощью метода визуализации данных высокой размерности *t-SNE*. Метод визуализации *t-SNE* - стохастическое вложение соседей с *t*-распределением, является техникой нелинейного снижения размерности, хорошо подходящей для вложения данных высокой размерности для визуализации в пространство низкой размерности.

Для построения полного патентного ландшафта используется модель LDA библиотеки gensim.

Для графиков кластеризации на координатной плоскости, ключевых слов каждого кластера, веса ключевых слов каждого кластера, а также сводной таблицы с распределением по темам используется модель LDA, предоставляемая библиотеками pySpark и sklearn.

Результаты

Программный модуль реализован на языках программирования Python и Scala и включает в себя 4 блока:

1) блок загрузки патентных архивов;
2) блок парсинга патентных массивов;
3) блок формирования набора данных патентных документов;
4) блок кластеризации патентов и визуализации патентного ландшафта.

Для парсинга патентных документов используется библиотека BeautifulSoup. Для кластеризации использованы библиотеки gensim, pySpark, sklearn. Визуализации моделей кластеризации основаны на библиотеках pyLDAvis и matplotlib. Разработанный модуль был апробирован на патентных массивах за 2019-2021 гг. На рис. 1 изображено распределение количества слов в документе. На рис. 2 изображена гистограмма с числом слов и их весом по срезу топиков.

Алгоритм машинного обучения для визуализации, представленный на рис. 3 и рис. 4, является техникой нелинейного снижения размерности, хорошо подходящей для вложения данных высокой размерности для визуализации в пространство низкой размерности. В частности, метод моделирует каждый объект высокой размерности двух- или трёхмерной точкой таким образом, что похожие объекты моделируются близко расположенными точками, а непохожие точки моделируются с большой вероятностью точками, расположенными далеко друг от друга.

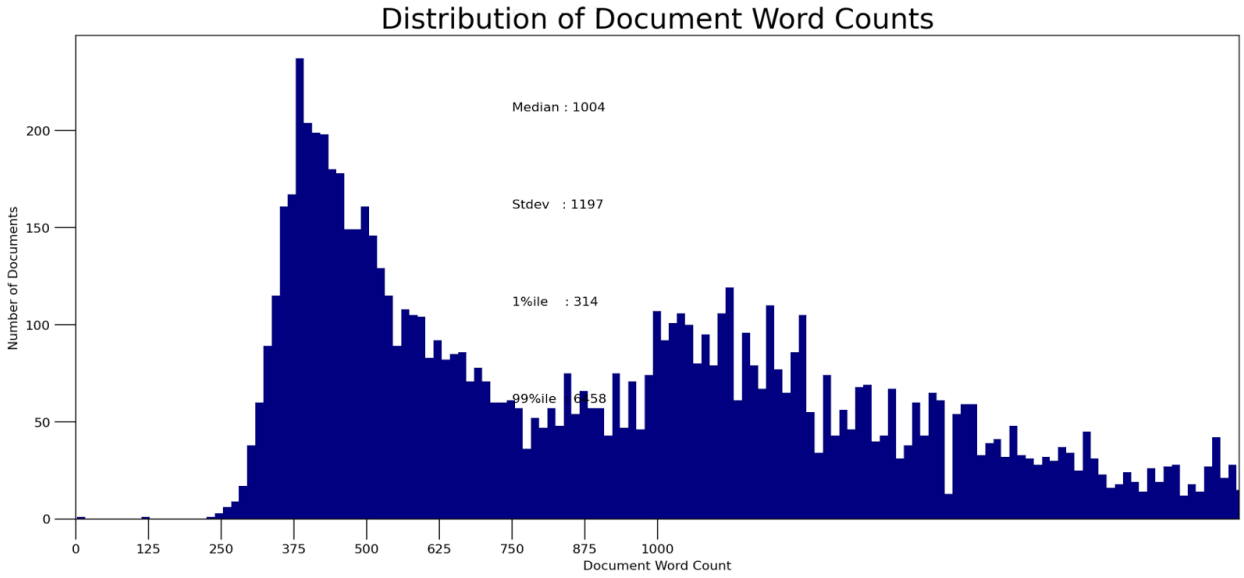

Рис. 1. – Распределение количества слов в документе

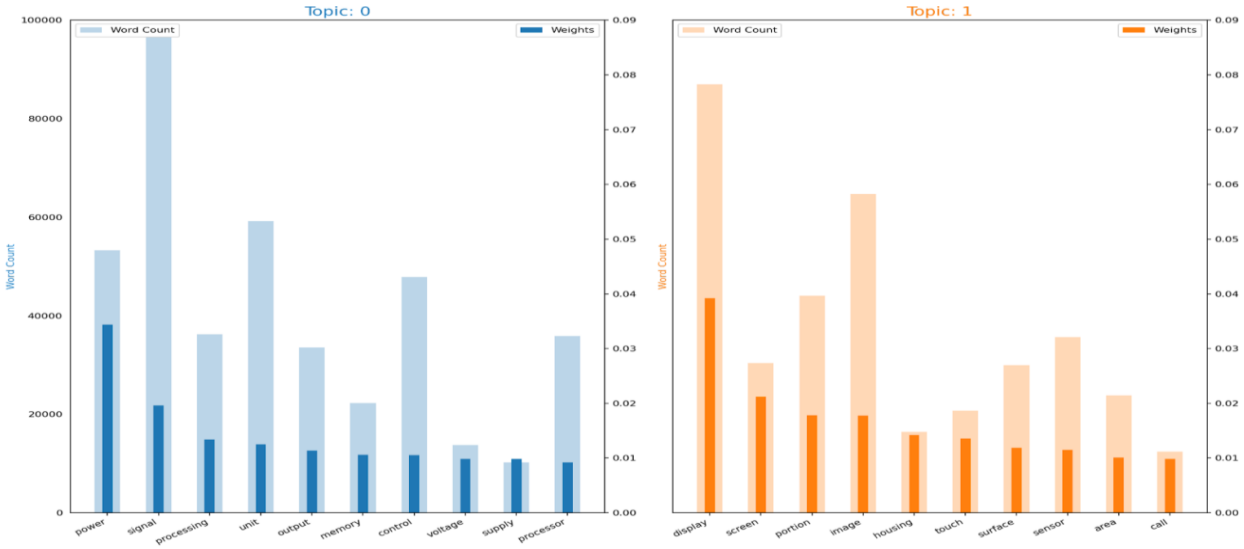

Рис. 2. – Число слов и из вес по срезу топиков

На рис. 3 представлено стохастическое вложение соседей с t-распределением на примере 2 категорий. На рис. 4 представлено стохастическое вложение соседей с t-распределением на примере 6 категорий.

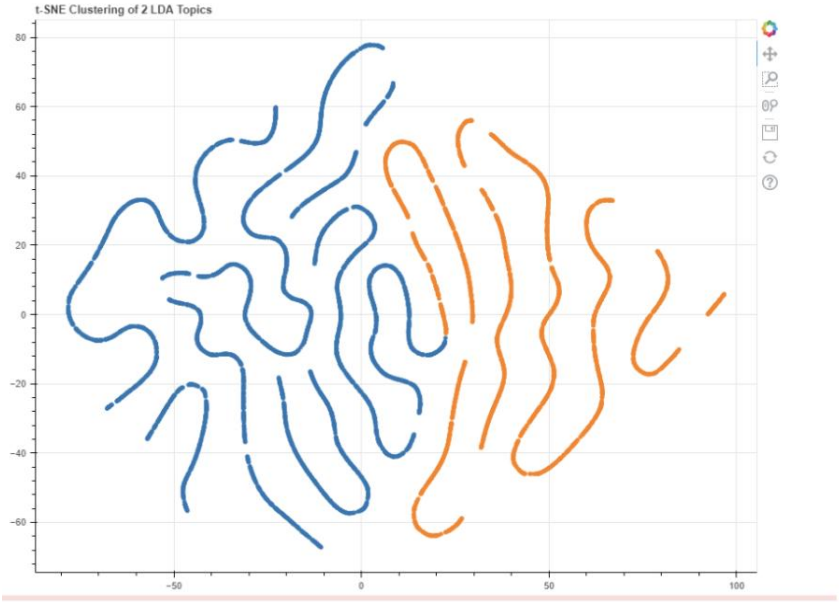

Рис. 3. – Стохастическое вложение соседей с t-распределением на примере 2 категорий

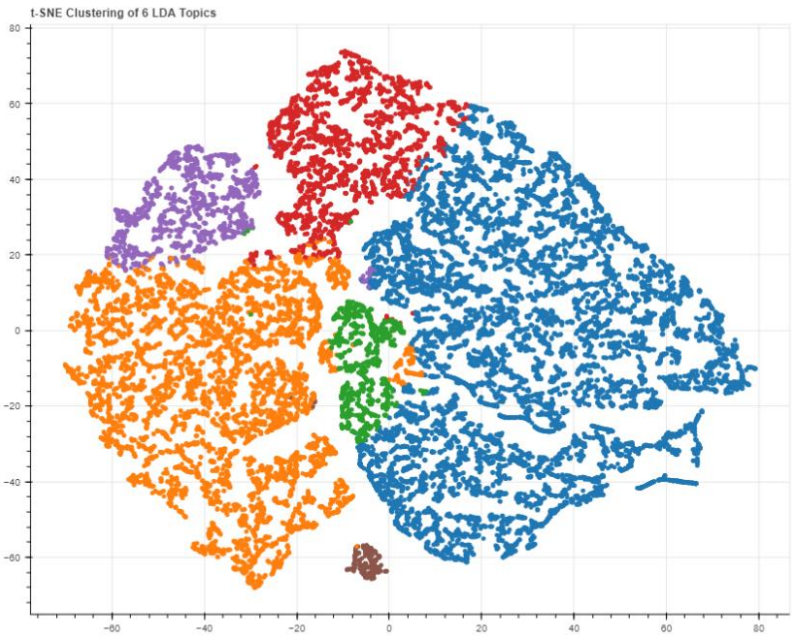

Рис. 4. – Стохастическое вложение соседей с t-распределением на примере 6 категорий.

На рис. 5 представлена интерактивная карта набора тем. Данная функциональность обеспечивается плагином pyLDAvis. В правой части карты изображены наиболее часто встречающиеся слова, входящие в тему. В левой части изображены темы в виде кругов. На рис. 6 изображено число слов по срезу топиков библиотеки sklearn.

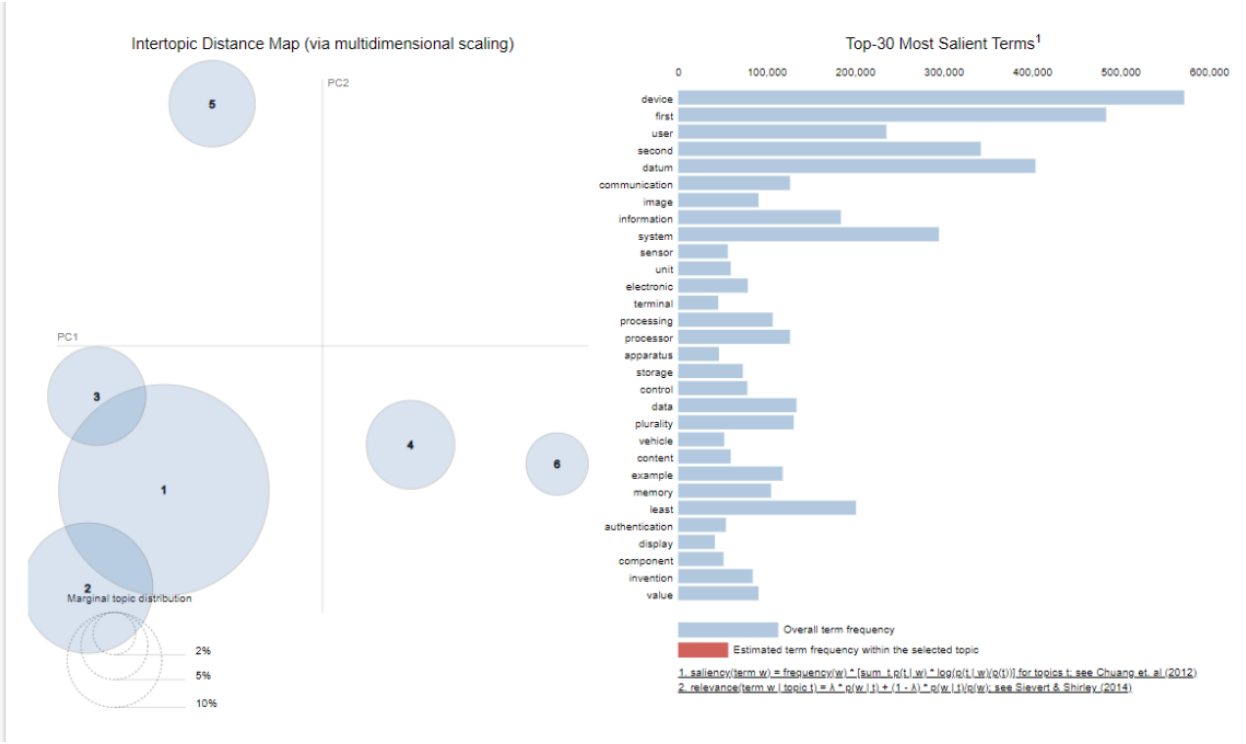

Рис. 5 – Интерактивная карта набора тем

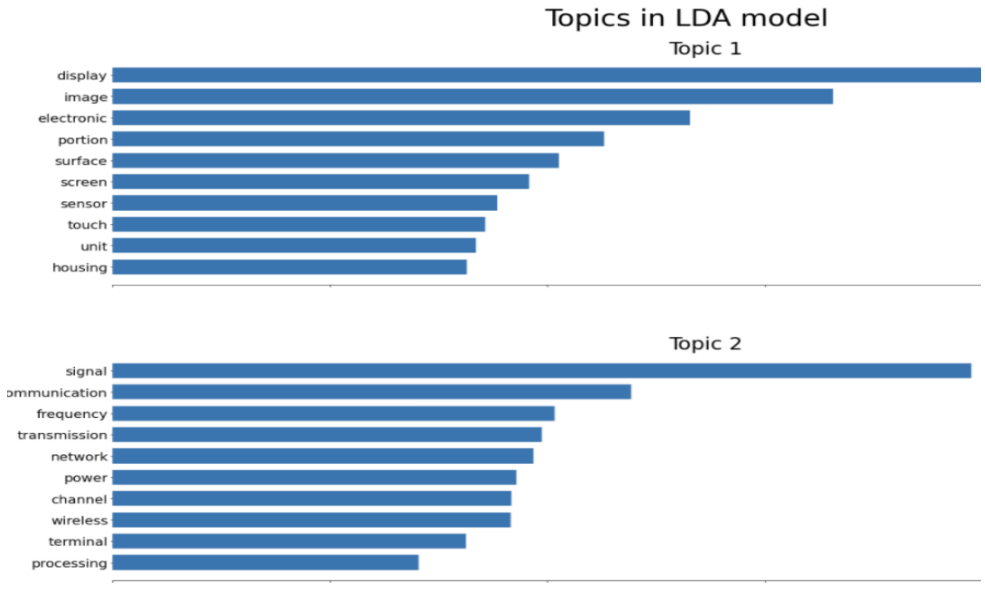

Рис. 6. – Число слов по срезу топиков библиотеки sklearn

Для модуля визуализации проведены 3 испытания. Цель испытаний - определить, правильно ли визуализируется патентный ландшафт с использованием различных входных данных, а также, насколько точна кластеризация патентов. Испытание №1 предназначено для того, чтобы выяснить, как объем набора данных влияет на работу программного модуля и точность кластеризации. Выбраны две очень похожие подкатегории патентов. Результаты испытания №1 представлены в таблице №1.

Таблица № 1

Испытание №1

Вид кластеризации	Точность кластеризации		
	Ожидаемый результат	Результат для 2000 патентов	Результат для 22000 патентов
Кластеризация Gensim	70%	50%	53%
Кластеризация PySpark	70%	50%	53%
Кластеризация sklearn	70%	67%	77%

Выводы на данном этапе заключаются в следующем:

1) PySpark и gensim показали недостаточную точность, необходимо выяснить причину;

2) кластеризация sklearn оказалась лучшим алгоритмом кластеризации, в дальнейшем необходимо выяснить причину;

3) возможно, что наборы данных недостаточны по объему для кластеризации, что повлияло на точность результата.

Испытание №2 предназначено для того, чтобы выяснить, как модель работает с двумя разными категориями патентов, которые имеют общую тему (и, следовательно, схожие ключевые слова). Результаты испытания №2 представлены в таблице №2.

Выводы на данном этапе заключаются в следующем:

1) Библиотеки gensim и PySpark показывают результаты лучше, чем в ходе испытания №1. Для набора данных из 10 000 патентов точность

увеличилась на 10-15% в сравнении с результатом для 22000 патентов испытания №1;

2) Sklearn более точен по сравнению с другими библиотеками;

3) При объеме 100 тыс. патентов на локальном компьютере PySpark не может выполнить кластеризацию. Для испытания требуется больше ресурсов.

Таблица № 2

Испытание №2

Вид кластеризации	Точность кластеризации		
	Ожидаемый результат	Результат для 10000 патентов	Результат для 100000 патентов
Кластеризация Gensim	70%	63%	69%
Кластеризация PySpark	70%	62%	
Кластеризация sklearn	70%	68%	74%

Результаты испытания №3 приведены в таблице №3.

Таблица № 3

Испытание №3

Вид кластеризации	Точность кластеризации
Кластеризация Gensim	80%
Кластеризация PySpark	84%
Кластеризация sklearn	80%

Был использован набор данных из 10 000 патентов. Кластеризация модели производилась по двум разным категориям патентов, не имеющих общих тем.

Выводы

В работе разработаны и описаны алгоритмы загрузки патентных архивов USPTO, парсинга патентных документов, кластеризации и визуализации патентного ландшафта. Алгоритмы реализованы в виде программного модуля. Программный модуль был апробирован на патентных документах USPTO за 2019-2021гг.

Использование данного разработанного программного модуля позволит облегчить процесс принятия стратегических решений в области патентования определённой области на уровне изобретателя или организации и повысить эффективность выявления новых тенденций и идей в области синтеза новых технических решений. Программный модуль может быть использован для исследования патентов, поиска взаимосвязей между патентами и построения визуальных представлений групп патентов по категориям.

Возможные способы улучшения системы: (а) добавить возможность сохранения модели LDA, обученной по определенным категориям патентов, для прогнозирования категории нового патентного документа; (б) улучшить качество предварительной обработки текста описания патента; (в) разрабатывать программное обеспечение на вычислительном кластере, чтобы полноценно использовать возможности фреймворка Apache Spark; (d) сохранять большой объем данных на вычислительном кластере с использованием технологии HDFS (распределенной файловой системы Hadoop).

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 22-21-20125, https://rscf.ru/project/22-21-20125/ и Администрации Волгоградской области.

Литература (References)

1. Korobkin D., Fomenkov S., Fomenkova M., Vayngolts I., Kravets A. Cyber-Physical Systems. Springer, Cham, 2021. pp. 161-172.

2. Korobkin D., Fomenkov S., Vereschak G., Kolesnikov S., Tolokin D., Kravets A. Cyber-Physical Systems. Springer, Cham, 2021. pp. 149-160.

3. Yang C., Huang C., Su J. Journal of informetrics. 2018. Т. 12. №. 1. pp. 271-286.

4. Kim S., Park I., Yoon B. Plos one. 2020. Т. 15. №. 2. P.0227930.

5. Guo J., Wang X., Li Q., Zhu D. Technological Forecasting and Social Change. 2016. Т. 105. pp. 27-40.

6. Wang X., Qiu P., Zhu D., Mitkova L., Lei M., Porter, A. Technological forecasting and social change. 2015. Т. 98. pp. 24-46.

7. Yang C., Zhu D., Wang X. International Journal of Computational Intelligence Systems. 2017. Т. 10. №. 1. p. 593.

8. Souili A., Cavallucci D., Rousselot F., Zanni C. Procedia Engineering. 2015. Т. 131. pp. 150-161.

9. De Prato G., Nepelski D., Piroli G. JRC Scientific and Policy Reports. EUR. 2015. Т. 27314. pp. 11-15.

10. Roh T., Jeong Y., Jang H., Yoon B. PloS one. 2019. Т. 14. №. 10. p. e0223404.

11. Feng L., Niu Y., Liu Z., Wang J., Zhang K. Sustainability. 2019. Т. 12. №. 1. p. 136.