Отчет о проделанной работе:

«Разработка программного модуля для совершенствования процесса иерархической кластеризации семантических связей между научно-техническими терминами, а также между документами на основе векторов семантико-синтаксических связей фраз и отдельных слов в документах»

Малашина Александра

Для решения данной задачи мною было написано 3 скрипта для решения следующих подзадач:

* terms\_occurence\_table.py

На входе получает таблицу терминов с id документа, в котором данный термин встречается. На выходе выдает таблицу совстречаемости слов в документах без повторов (AB = BA).

Для использования программы по построению таблицы совстречаемости слов (terms\_occurence\_table.py) необходимо наличие следующих библиотек:

1. csv

необходима для работы с csv файлами

2. time

необходима для контроля за временем работы программы

* terms\_distance\_table.py

На вход получает таблицу совстречаемости терминов, сделанную на предыдущем этапе, на выходе выдает таблицу близости терминов по четырем метрикам ('cosine', 'jaccard', 'dice', 'cityblock').

Для использования программы по расчету близости векторов слов (terms\_distance\_table.py) необходимо наличие следующих библиотек:

1. csv

необходима для работы с csv файлами

2. time

необходима для контроля за временем работы программы

3. scipy.spatial.distance

необходима для расчета близости векторов

* hierarchical\_clustering.py

На вход получает таблицу терминов с id документа, в котором данный термин встречается. Производит все итерации из предыдущих скриптов и на выходе выдает дендрограмму, построенную по матрице близости и отображающую взаимные связи между терминами заданного множества.

Для использования программы по построению иерархической кластеризации (hierarchical\_clustering.py) необходимо наличие следующих библиотек:

1. time

необходима для контроля за временем работы программы

2. matplotlib.pyplot

необходима для построения графика иерархической кластеризации

3. scipy.cluster

необходима для построения иерархической кластеризации

4. matplotlib.backends.backend\_pdf

необходима для сохранения построенного графика в pdf формате

5. argparse

необходима для работы из командной строки

скрипты:

1. terms\_occurence\_table

необходим для предворительного создания таблицы совстречаемости слов

2. terms\_distance\_table

необходим для предворительного создания таблицы близости слов

Программа запускается через командную строку со следующими параметрами:

Ключи для запуска программы через командную строку:

-h, --help show this help message and exit

-d DISTANCE, --distance DISTANCE

Choose distance metric: 'cosine', 'jaccard', 'dice', 'cityblock'

-m METHOD, --method METHOD

Choose clustering method: 'single', 'complete','average', 'weighted'