План проведения численного эксперимента

поиска семантических дубликатов в новостных сообщениях

Введение

Задача *поиска семантических дубликатов* сообщений сводится к поиску сообщений близких по смыслу и отличается от *задачи поиска нечетких дубликатов.* Отличие заключается в том, что при поиске нечетких дубликатов тексты лексически сильно похожи, т.е. у двух текстов будет совпадать большое количество слов, словосочетаний и предложений. При поиске семантических дубликатов будет совпадать ограниченное количество слов.

Поэтому необходимо провести численный эксперимент, который позволит установить параметры, при которых два текста можно считать схожими по смыслу.

Исходные данные

Тексты новостных сообщений поступают в систему СМАД из *rss*-лент. Особенность таких текстов заключается в их длине. Средняя длина сообщений составляет *37* слов (рис. 1), среднее количество уникальных слов – *32* (рис. 1). В табл. 1 представлены подробные характеристики сообщений по часам.

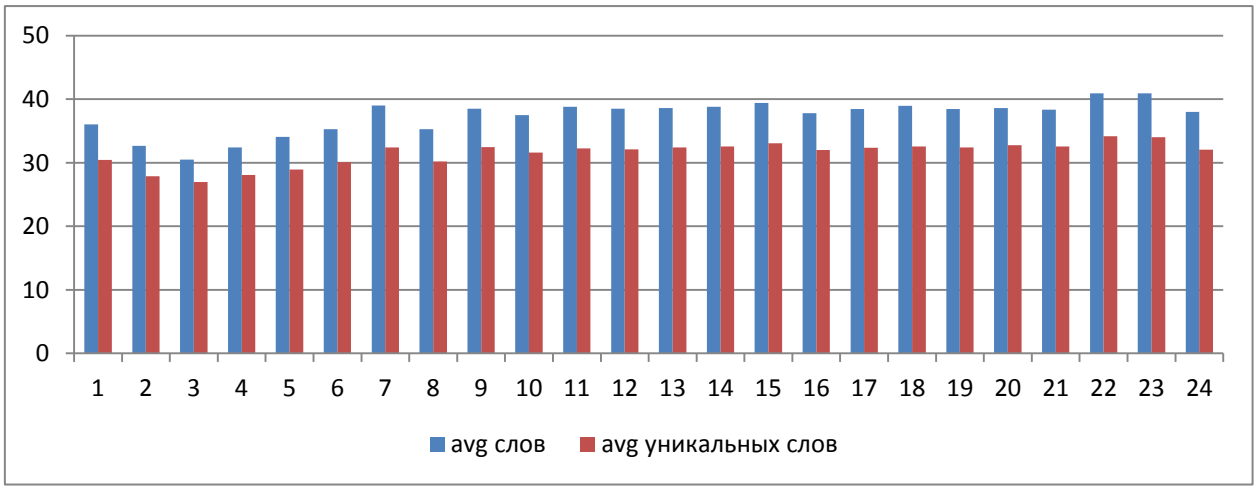

рис. 1

Таблица 1.

Час	**Среднее кол-во текстов**	**Среднее кол-во слов за час**	**Среднее кол-во уникальных слов за час**	**Процент уникальных слов**
00 – 01	457	36	30	84%
01 – 02	245	33	28	85%
02 – 03	322	30	27	88%
03 – 04	248	32	28	87%
04 – 05	293	34	29	85%
05 – 06	368	35	30	85%
06 – 07	513	39	32	83%
07 – 08	656	35	30	86%
08 – 09	959	39	32	84%
09 – 10	1199	37	32	84%
10 – 11	1392	39	32	83%
11 – 12	1463	38	32	83%
12 – 13	1440	39	32	84%
13 – 14	1381	39	33	84%
14 – 15	1359	39	33	84%
15 – 16	1317	38	32	85%
16 – 17	1262	38	32	84%
17 – 18	1133	39	33	84%
18 – 19	907	38	32	84%
19 – 20	734	39	33	85%
20 – 21	632	38	33	85%
21 – 22	489	41	34	84%
22 – 23	408	41	34	83%
23 – 00	325	38	32	84%
Среднее	**813**	**37**	**32**	**85%**

Для постановки эксперимента будет использоваться коллекция новостных сообщений с сервиса Яндекс.Новости. Каждое сообщение содержит заголовок, тело (основную часть) и информацию о принадлежности к тематическому кластеру. Принадлежность к тематическому кластеру будет идентифицировать два случайно взятых сообщения как «дубликаты» или «не дубликаты». Пример данных для постановки эксперимента представлен в табл. 2. Сама тестовая коллекция текстов содержится в файле *test.csv.*

Таблица 2.

№ п/п	Заголовок	Основная часть	Кластер
1	Набиуллина назвала близким к обоснованному текущий курс рубля	Ранее в среду курс доллара на Московской бирже подскочил больше чем на 2 руб. и достиг наивысшего показателя со времен деноминации 1998 года.	1

2	Набиуллина не видит риска для финансовой стабильности	*Курс рубля* близок к *фундаментальным уровням*, риска для финансовой стабильности нет, считает *глава российского Центробанка Эльвира Набиуллина*.	1
3	Набиуллина прокомментировала курс рубля	*Глава Банка России Эльвира Набиуллина* полагает, что текущий *курс рубля* близок к своему *фундаментально* обоснованному *курсу*, передает ТАСС. "Да, сейчас близок", - заявила Набиуллина, не уточнив при этом, куда должен пойти курс рубля, чтобы стать фундаментально обоснованным.	1
4	Песков: готовая нарушить закон внесистемная оппозиция не способствует стабильности России	Песков не стал комментировать высказывания главы Чечни Рамзана Кадырова о внесистемной оппозиции, он призвал внимательнее вчитаться в слова чеченского руководителя.	2
5	Гамбургский балет открывает гастроли в Большом театре	В Большом театре на Исторической сцене гастроли откроет сегодня, 20 января, Гамбургский балет. Постановки труппы россияне смогут увидеть до 24 января включительно. Руководит Гамбургским балетом на протяжении 40 лет Джон Ноймайер, известный хореограф и вдохновитель труппы.	3
6	Урин: приглашение зарубежных артистов в ГАБТ укрепит культурные связи	Мировая премьера состоялась в 1989 году на сцене Гамбургского театра, а через полгода спектакль был показан в Большом в Москве.	3

Описание эксперимента

Визуальный анализ текстов №2 и №3, представленных в табл. 2, позволяет сформулировать следующее продукционное правило:

*«**Если*** у двух текстов явно повторяются имена собственные (Эльвира Набиуллина) *и* явно повторяется ≥*3* других слов (курс, рубль, глава) *и* имеется ≥*2* неявных повторов (российского Центробанка *vs* Банка России, фундаментальным *vs* фундаментально), ***То*** такие тексты можно считать ***дубликатами»***. Сразу хотелось бы оговориться, что представленное правило не претендует на полноту и правильность.

Так вот суть эксперимента заключается в том, чтобы *сформулировать полный перечень подобных продукционных правил из анализа тестовой выборки*. Для этого предлагается провести частотно-морфологический и семантический анализ пар текстовых сообщений из тестовой выборки.

Введем следующие понятия

Явный повтор. Это точное совпадение слов или словосочетаний в нормальной форме. Например, *«белый снег» = «белым снегом».* Для явно повторяющихся слов следует выделять граммемы: имя (имя собственное, ФИО, географические названия, сокращения), существительное, прилагательное, числительное и т.д. К явным повторам относятся опечатки. Так *«Опечатка» = «очепятка»* или *«Новый Год» = «Новыйгод»*.

Неявный повтор. К неявным повторам относятся следующие виды совпадений: 1) совпадение, при котором слово меняет часть речи, но лемма сохраняется *«Хабаровск» = «Хабаровский»* – общая лемма *Хабаровск, «бойкотируют» = объявили «бойкот»* – общая лемма *бойкот*; 2) полная либо краткая форма прилагательных *«Веселый» = «весел», «красивый» = «красив»*.

Здесь очень важно понимать, что *«киностудия» ≠ «фотостудия»,* несмотря на то, что у этих слов будет один и тот же корень *студия.*

Семантический повтор. Это схожесть слов или словосочетаний по смыслу или по близости употребления. Например,

по смыслу: *«негр» = «афроамериканец» = «темнокожий» = «чернокожий». «Российский Центробанк» = «Банк России».*

по близости употребления: *«Арабская весна» = «твиттер-революция». «Холодильное оборудование» = «монтировать». «Россиянка» = «Черкесска» = «Татарка»* и т.д.

При сравнении двух текстов будет производиться подсчет повторяющихся частей. Для примера вернемся к сравнению текстов №2 и №3 из табл. 2. В этих текстах 2 раза явно повторяется граммема «имя собственное» *(имя),* 3 раза явно повторяется граммема «существительное» *(S)*, 1 неявный повтор (*«фундаментальным» = «фундаментально»*), 2 семантических повтора (*«уровням» = «курсу», «российского Центробанка» = «Банка России»*).

Подобные сравнения необходимо провести для всех пар, которые получаются в обучающей коллекции текстов *test.csv* и занести в табл. 3. Коли-

чество пар будет равняться C_n^2, где n – объем тестовой коллекции текстов *test.csv*. Наименование всех граммем для табл. 3 приведено на сайте ПО *mystem* https://tech.yandex.ru/mystem/doc/grammemes-values-docpage/

Таблица 3.

№ п/п	Явный повтор					Неявный повтор	Семантический повтор	Дубликат
	имя	*A*	*ADV*	*...*	*V*			
1								
2								
3								
…								
C_n^2								
	X							*Y*

Примечание: здесь граммема *«имя»* включает в себя географические названия, имена собственные, ФИО, сокращения.

Каждая строка табл. 3 представляет собой результат сравнения двух текстов из тестовой коллекции текстов. В каждую строку табл. 3 заносится следующая информация: количество совпавших граммем при явном повторе, количество неявных и семантических повторов. Эти данные образуют вектор входных независимых (объясняющих) переменных *X*, а в качестве выходной – объясняемой переменной *Y* выступает идентификатор дубликата, т.е. если два текста (как в примере с Эльвирой Набиулиной) принадлежат одному тематическому кластеру (последний столбец табл. 2), то они являются дубликатами.

Заполнив такую таблицу можно:

1. Выявить наиболее значимые переменные *(x_1, x_2, ...)* путем факторного анализа или другим методом.

2. Разработать математическую модель, которая будет предсказывать по входному вектору *X=(x_1, x_2, ...)* значение выходной переменной *Y*. Для этого можно: 1) построить дерево решений *(TreeDecision)*. Каждая ветвь дерева

представляет собой продукционное правило вида *«Если A, то B»;* 2) используя алгоритмы *Apriori* или *FP-growth* построить ассоциативные правила вида *«Если A, то B»;* 3) использовать *SVM-классификатор;* 4) использовать различные методы кластеризации и др.

Проблемы, с которыми придется столкнуться при постановке эксперимента:

1. Выявление граммем. Для этого, на мой взгляд, необходимо использовать *Томита-парсер (https://tech.yandex.ru/tomita/)* и *mystem (https://tech.yandex.ru/mystem/)* в чистом виде и писать собственные грамматики, например, для поиска имен собственных. На эти инструменты обратить особое внимание! Обратите внимание на параметр (-d) для *mystem,* этот параметр снимает омонимию для слов. Также обратите внимание на то, что *mystem* позволяет определять имена собственные, географические названия и т.д., но иногда делает это с ошибками, поэтому необходимо продумать собственные грамматики для *Томита-парсер.*

Те слова, которые отсутствуют в словаре – не идентифицируются никаким образом. Так, например, с точки зрения явных повторений слова *«Центробанка»* и *«Центробанком»* являются повторами, но с точки зрения *mystem* – это разно идентифицируемые слова*;*

Обращаю внимание на настройку параметров mystem!!!

2. Выявление орфографических ошибок и опечаток. Для поиска орфографических ошибок и опечаток рекомендуется использовать *Спеллер (https://tech.yandex.ru/speller/).*

3. Поиск неявных повторов. Для этого можно использовать *mystem (https://tech.yandex.ru/mystem/)* и *расстояние Левенштейна*.

4. Поиск семантических повторов. Скорей всего здесь необходимо очень мудро использовать *W2V.* Думается, что есть еще какие-то графовые или сетевые методы поиска семантических повторов.