МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
КАФЕДРА МО ЭВМ

ОТЧЕТ

по научно-исследовательской работе

Тема: разработка автоматизированного метода оценки соответствия между речью и презентацией докладчика

Студентка гр. 6304 Тарасова А.А.

Преподаватель Заславский М.М.

Санкт-Петербург
2021

ЗАДАНИЕ

НА НАУЧНО-ИССЛЕДОВАТЕЛЬСКУЮ РАБОТУ

Студентка Тарасова А.А.

Группа 6304

Тема работы: Разработка автоматизированного метода оценки соответствия между речью и презентацией докладчика

Исходные данные:
Транскрипции тренировок и данные с презентаций из Тренажера публичных выступлений

Содержание пояснительной записки:
«Содержание», «Введение», «Постановка задачи», «Результаты работы в осеннем семестре», «План на весенний семестр», «Список использованных источников»

Предполагаемый объем пояснительной записки:
Не менее 10 страниц.

Дата выдачи задания: 21.12.2020

Дата сдачи реферата: 27.12.2021

Дата защиты реферата: 28.12.2021

Студентка _____ Тарасова А.А.

Преподаватель _____ Заславский М.М.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Оценка качества публичного выступления всегда субъективна. Для подготовки к выступлению и повышения его качества, как правило, необходимы человеческие ресурсы: люди, готовые прослушать доклад и дать рекомендации. В целях экономии человеческих ресурсов актуально построение автоматизированной системы, способной выдавать не только числовую оценку по некой шкале, но и базовые рекомендации, ведущие к повышению качества доклада.

Система автоматизированной оценки докладов была разработана кафедрой МО ЭВМ, она носит название “Тренажер публичных выступлений” [1]. В рамках этой системы на данный момент оценка выдается на основе соответствия общей длительности выступления регламенту, скорости речи и наличия слов-паразитов. Для развития этой системы была поставлена задача разработки новых критериев.

Критериев для оценки качества выступления можно выделить много, но почти все будут выдавать субъективную оценку. Целью данной работы является построение алгоритма, способного дать оценку соответствия произнесенной докладчиком речи плану, обозначенному в его презентации. Выявление отклонений может свидетельствовать либо о неуверенности докладчика в заявленном материале, либо о необходимости редактирования и актуализации сопроводительного материала. В таком случае для оценки и выдачи рекомендаций можно опираться на две составляющие: общее соответствие по всему докладу и послайдовая оценка.

Данная задача относится к области NLP [2] - обработке естественных языков, для ее решения использованы инструменты обработки текстов в целом и отдельных слов в частности: выделение ключевых слов, составление словарей, лемматизация, токенизация, стемминг.

1. ПОСТАНОВКА ЗАДАЧИ

Так как цель научно-исследовательской работы – построение автоматизированного метода оценки соответствия между речью и презентацией докладчика – поставлена в рамках развития существующей системы, Тренажера публичных выступлений [1], итоговый критерий должен быть реализован с использованием open source технологий, работать быстро (время оценки не должно превышать 3 секунд). Также необходимо избегать значительных затрат по памяти и гарантировать стабильную работу критерия, по возможности использовать уже введенные в систему инструменты.

В рамках данной работы необходимо решить задачу обработки естественного языка для определения содержания речи, опираясь на извлечение ключевых слов и учитывая специфику работы с устной речью, и оценки сходства транскрипции речи и содержимого презентации.

Составление такого критерия состоит из следующих этапов:

- Рассмотрение механизма извлечения ключевых слов и его реализация на языке Python
- Адаптация под извлечение КС для распознанной устной речи
- Создание механизма сравнения текстов по ключевым словам с учетом специфики решаемой задачи – расширения системы тренировки публичных выступлений
- Добавление критерия в существующую систему
- Тестирование критерия на новых тренировках
- Подбор оптимальных значений параметров критерия

2. РЕЗУЛЬТАТЫ РАБОТЫ В ОСЕННЕМ СЕМЕСТРЕ

2.1. План, обозначенный по итогам весеннего семестра

В предыдущем семестре был построен алгоритм извлечения ключевых слов на базе tf-idf метрики с использованием таких технологий, как nltk [3] и pymorphy2 [4], которые уже использовались в других компонентах тренажера публичных выступлений.

1. Внедрение в систему тренировки публичных выступлений разработанного критерия, его адаптация для работы с сравнением по отдельным слайдам.
2. Тестирование корректности работы на реальных данных: распознанных текстах и загруженных презентациях, в том числе на «неудачных» примерах работы распознавателя, зашумленных файлах.
3. Установление оптимальных пороговых значений метрики tf-idf или
4. количества ключевых слов для сравнения.
5. Отображение полученного значения критерия в оценку выступления с учетом, оценка необходимости полного вхождения ключевых слов презентации в речь докладчика.
6. Загрузка текстов в корпус из базы данных, используемой в существующей системе.

2.2. Внедрение критерия в систему тренировки публичных выступлений

На данный момент критерий имеет 2 различных реализации, которые будут тестироваться на новых тренировках со специальным допущением ошибок. Такой подход позволит принять субъективное решение о том, какой из построенных алгоритмов выдает наиболее близкий к предполагаемому результат.

Описание алгоритма 1:

1. Производится предобработка текста, текст парсится на стемы отдельных слов
2. Строится множество ключевых слов со слайда – kw_{slide}
3. Строится множество ключевых слов из транскрипции речи за время, когда данный слайд был открыт – kw_{speech}
4. Строится множество слов, которые были распознаны в транскрипцию слайда, но не вошли в список ключевых – $speech$
5. Определяется пересечение множеств kw_{slide} и kw_{speech} – называем это множество совпадением – $coinc$
6. Определяется пересечение множеств kw_{slide} и $speech$ – называем это множество различием, diff
7. Результирующая оценка:

$$\frac{len(coinc)}{len(coinc\ +\ diff)}$$

Данный алгоритм считает, что если ключевое слово презентации не было распознано в речи (использование стем позволяет снизить влияние ошибок распознавателя), то оно распознаватель ошибся. Иначе говоря, он искусственно завышает оценку, которая скорее всего искусственно снижается ошибками распознавателя.

В силу этого был разработан второй алгоритм.

Описание алгоритма 2

1. Производится предобработка текста, текст парсится на стемы отдельных слов
2. Строится множество ключевых слов со слайда – kw_{slide}
3. Строится множество всех слов из транскрипции – $speech$
4. Определяется пересечение множеств kw_{slide} и $speech$ – называем это совпадением $coinc$
5. Результирующая оценка:

$$\frac{len(coinc)}{len(kw_{slide})}$$

Оценка, выдаваемая этим алгоритмом, более честная. Но в результате получается довольно низкое цифровое значение, которое далеко не всегда оправдано, оценка остается заниженной из-за специфики работы распознавателя.

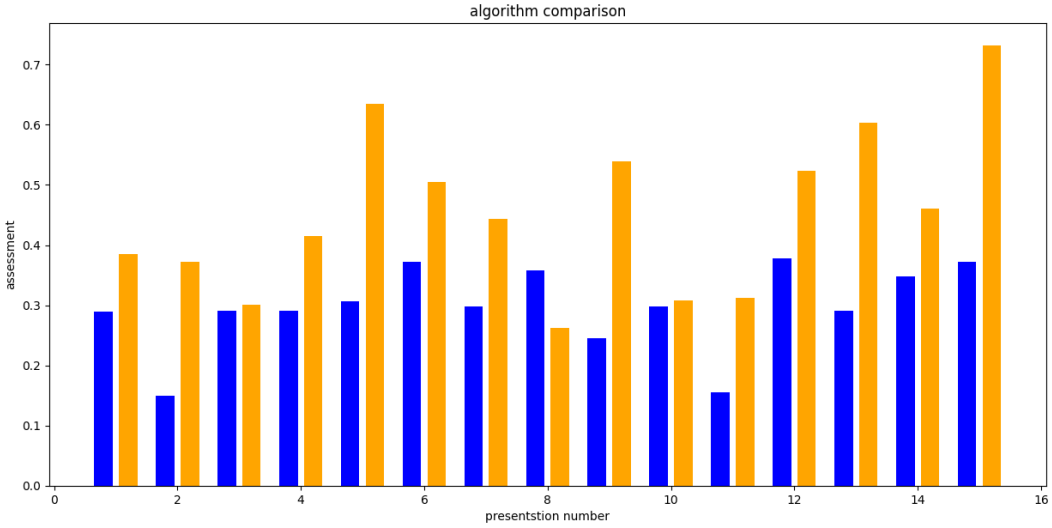

Рисунок 1. Сравнение оценок, выданных двумя алгоритмами, на 15 презентациях

Также результаты послайдовой оценки приведены на рисунках ниже (один график – одна презентация, столбцам соответствуют слайды).

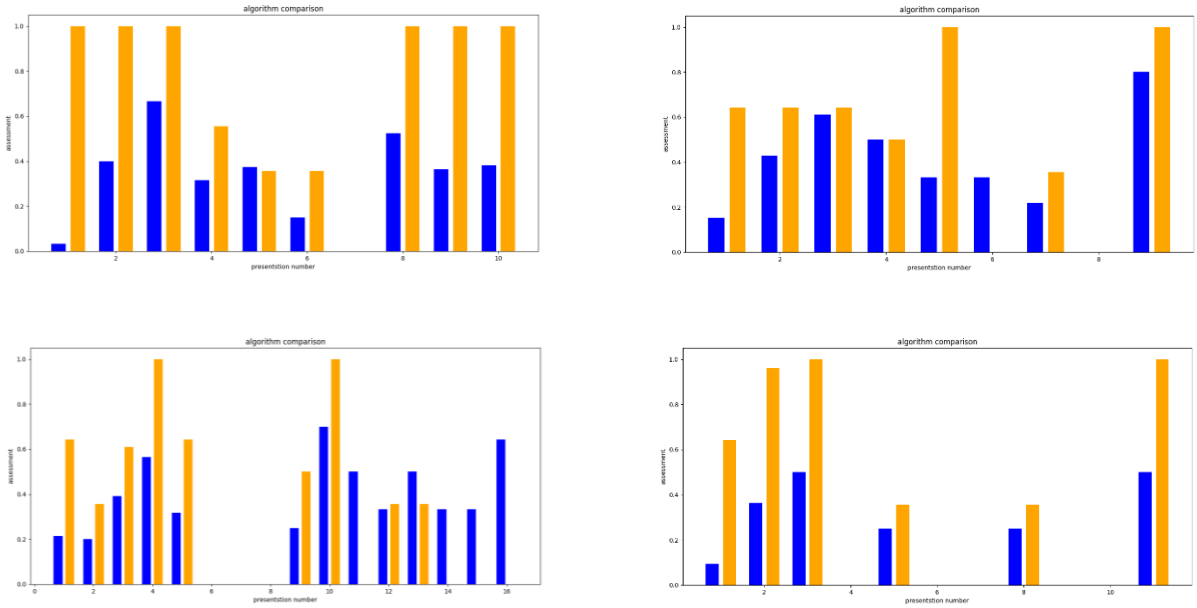

Рисунок 2. Сравнение оценок по слайдам

Обратимся к последнему рисунку, на котором видно очень низкую суммарную оценку (не считая нескольких единиц). Анализ этой тренировки показал, что речь докладчика была очень нечеткой, и в основном идейно

ключевые слова слайда в речи произнесены были, но распознать их может только подготовленный человек. Видим, что оранжевые столбцы иногда показывают высокую оценку совпадения там, где более честный алгоритм имеет вдвое меньшую оценку.

На основе анализа нескольких тренировок была построена следующая идея. Если разница между оценками двух алгоритмов значительна (обозначив некоторый порог этого), то возможно, что транскрипция содержит большое количество ошибок, в таком случае слайд принимать во внимание не стоит, либо идея на слайде отличается от произнесенной. Для таких слайдов следует опираться на df-метрику: если в тематическом словаре (специфичном для предметной области, в которой идут тренировки) не содержится распознанных слов, то проблема находится на стороне распознавателя. В таком случае принимать оценку данного слайда в расчете оценки выступления не стоит.

Помимо задачи, поставленной в рамках предыдущего семестра – выполнения сравнения по каждому слайду отдельно (каждый слайд получает оценку), выполняется также общая оценка речи. Иначе говоря, оценивается соответствие речи и текста как по основной линии доклада, так и по содержимому каждого слайда.

2.3. Пороговые значения

Критерий принимает на вход 4 параметра: пороговый уровень метрики tf-idf и предельное количество ключевых слов для презентации и для транскрипции (для транскрипции это актуально не только для 1 алгоритма, но и для сравнения идейной линии презентации и речи в целом).

Опираясь на ошибки распознавателя, необходимо проводить сравнение на основе ключевых слов презентации (которые точно не содержат ошибок, хотя здесь не принимается во внимание возможное наличие орфографических ошибок, если они содержатся в корне - принимаем орфографию близкой к идеальной хотя бы в рамках стем). Второй пункт, необходимый для снижения доли ошибок – добавление во множество ключевых большего количества слов

для речи, чем для презентации. Тем самым обходится ситуация, когда ошибка распознавателя становится ключевым словом и при равных мощностях множеств приводит к занижению оценки.

Результаты сравнения результатов для различных значений метрик представлены в таблицах в приложениях А и Б.

2.4. Тексты для вычисления df-метрики

Задача 5, загрузка текстов, была изменена: предположение о рациональности работы с корпусом текстов (подключение базы данных к работе критерия для подсчета df-метрики) привело бы к значительным расходам памяти: постоянное обновление базы с ростом числа проведенных тренировок, во-первых, привело бы к накладным расходам на загрузку и расчет метрики при запуске сервера, а во-вторых, затруднило бы контроль над переполнением памяти и могло бы привести к нестабильной работе критерия, а значит, и системы в целом.

Задача 5 была видоизменена на составление словаря, адаптированного под область научных интересов, в которой происходит большая часть тренировок. Данная задача была решена на основе обработки презентаций и транскрипций: были подсчитаны наиболее частоупотребимые слова (точнее, их стеммы – значимые части слова), далее они вручную были распределены на 2 словаря: первый – общеупотребимые слова, вес которых в сравнении текстов должен быть снижен, второй – слова, имеющие повышенный вес в силу значимости в основной части проанализированных текстов.

Задача, не обозначенная в плане, но также решенная в рамках осеннего семестра – различие весов слайдов на основе их заголовков.

2.5. Веса слайдов

Интуитивно понятно, что не все слайды презентации равнозначны. Поэтому стоит различать слайды, имеющие вводный характер, и слайды, которые определяют суть всего доклада.

Анализ 104 тренировок, проведенных в системе, выявил наиболее частоиспользуемые в заголовках слова.

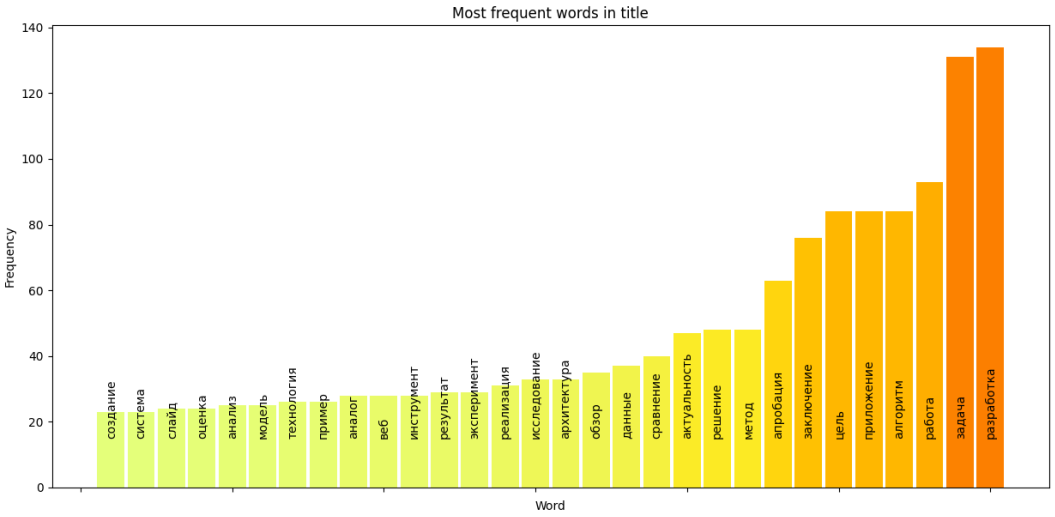

Рисунок 3. Наиболее частоупотребимые названия

Далее они вручную были распределены на 3 группы:

- наиболее значимые и информативные слайды:

  'задача', 'приложение', 'цель', 'актуальность', 'обзор', 'решение', 'результат'

- слайды, вес которых должен быть снижен:

  'пример', 'модель', 'данные', 'технология', 'спасибо', 'внимание', 'команда', 'история', 'развитие'

- слайды, требующие дополнительного анализа для выявления веса:

  'результат', 'эксперимент', 'апробация', 'метод' .

В третьей группе находятся слайды, содержимое которых важно с точки зрения презентации результатов работы. Однако такие слайды часто сопровождаются иллюстративным материалом, содержат англицизмы. Проблема англицизмов заключается в том, что они не могут быть распознаны с помощью vosk [5] – инструмента распознавания речи, который используется в тренажере (он поддерживает как русский, так и английский язык, но может работать только с одним из них в зависимости от настроек и не способен отличать языки).

3. ОПИСАНИЕ ПРЕДПОЛАГАЕМОГО МЕТОДА РЕШЕНИЯ

Описанные в предыдущей главе решения составляют 2 критерия, которые будут опробованы на тренировках в системе, в следствие чего будет выбран только один из них.

В общем виде алгоритм следующий:

1) Слайд и транскрипция проверяются на пустоту
2) Инициализируются массивы: один для весов слайдов, другой для оценок
3) Цикл по содержательным слайдам:
   A) текст транскрипции и слайд проходят предобработку, разделяются на токены и фильтруются
   B) из слайда и транскрипции извлекаются ключевые слова
   C) по заголовку определяется вес слайда
   D) применяется один из описанных алгоритмов, слайд получает оценку содержания
   E) вес слайда и его оценка записываются в массив
4) По заполненным массивам получаем общую оценку презентации
5) Определяем ключевые для всей презентации слова и ключевые для всей транскрипции
6) Применяем алгоритм сравнения ключевых слов

Результаты работы находятся в двух ветках репозитория тренажера публичных выступлений

4. ПЛАН НА ВЕСЕННИЙ СЕМЕСТР

1. Выбор алгоритма на основе проведенных (в том числе намеренно искаженных) тренировок
2. Дополнение словаря общеупотребимых слов и словаря специальных терминов
3. Проверка гипотезы о замене извлечения ключевых слов на работу со словарем специальных терминов
4. Отладка работы критерия внутри Тренажера публичных выступлений
5. Завершение работы над статьей и подготовка доклада

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Репозиторий проекта Web Speech Trainer // github.com. https://github.com/OSLL/web_speech_trainer
2. Recent Trends in Deep Learning Based Natural Language Processing / Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria. arXiv:1708.02709 [cs], 2018. 32 с.
3. Описание библиотеки nltk на официальном сайте // URL: https://www.nltk.org/
4. Описание синтаксического анализатора pymorphy2 // pymorphy2.readthedocs.io. URL: https://pymorphy2.readthedocs.io/en/latest/
5. Инструмент распознавания речи Vosk // alphacephei.com. URL: https://alphacephei.com/vosk/
6. Ветки репозитория Тренажера публичных выступлений с реализацией критерия // github.com. URL: https://github.com/OSLL/web_speech_trainer/tree/keywords_extraction https://github.com/OSLL/web_speech_trainer/tree/speech_and_presentation_comparator
7. Ссылка на черновик статьи для выступления на Научно-техническом семинаре кафедры МО ЭВМ // URL: https://docs.google.com/document/d/1FMfP2Na1UVQa-EHsYIfmZzOMXURRSX-mIsCmkbPwWJg/edit?usp=sharing

ПРИЛОЖЕНИЕ А. ТАБЛИЦА СРАВНЕНИЯ КОЛИЧЕСТВА КЛЮЧЕВЫХ СЛОВ

count_slide	count_speech	algorithm_1	algorithm_1	max_1	max_2
5	6	0,46	0,24	0,70	0,46
5	7	0,46	0,28	0,70	0,56
5	8	0,46	0,31	0,70	0,58
5	9	0,46	0,33	0,70	0,60
5	10	0,46	0,34	0,70	0,61
5	11	0,46	0,36	0,70	0,61
5	12	0,46	0,38	0,70	0,61
6	7	0,47	0,28	0,76	0,57
6	8	0,47	0,32	0,76	0,59
6	9	0,47	0,33	0,76	0,60
6	10	0,47	0,34	0,76	0,60
6	11	0,47	0,36	0,76	0,60
6	12	0,47	0,38	0,76	0,61
7	8	0,48	0,33	0,74	0,58
7	9	0,48	0,34	0,74	0,59
7	10	0,48	0,35	0,74	0,66
7	11	0,48	0,37	0,74	0,66
7	12	0,48	0,39	0,74	0,66
8	9	0,48	0,36	0,72	0,60
8	10	0,48	0,37	0,72	0,60
8	11	0,48	0,39	0,72	0,62
8	12	0,48	0,41	0,72	1
9	10	0,48	0,39	0,72	0,67
9	11	0,48	0,41	0,72	0,67
9	12	0,48	0,42	0,72	0,68
10	11	0,48	0,42	0,68	0,64
10	12	0,48	0,43	0,68	0,66
11	12	0,47	0,42	0,65	0,63

ПРИЛОЖЕНИЕ Б. ТАБЛИЦА СРАВНЕНИЯ КОЛИЧЕСТВА КЛЮЧЕВЫХ СЛОВ

level_slide	level_speech	algorithm_1	algorithm_1	max_1	max_2
0,1	0,2	0,43	0,77	0,56	0,94
0,1	0,3	0,43	0,69	0,56	0,93
0,1	0,4	0,43	0,59	0,56	0,78
0,1	0,5	0,43	0,34	0,56	0,51
0,1	0,6	0,43	0,34	0,56	0,51
0,1	0,7	0,43	0,31	0,56	0,46
0,2	0,3	0,43	0,70	0,65	0,88
0,2	0,4	0,43	0,61	0,65	0,78
0,2	0,5	0,43	0,37	0,65	0,55
0,2	0,6	0,43	0,36	0,65	0,55
0,2	0,7	0,43	0,32	0,65	0,47
0,3	0,3	0,45	0,67	0,66	0,83
0,3	0,4	0,45	0,58	0,66	0,77
0,3	0,5	0,45	0,35	0,66	0,57
0,3	0,6	0,45	0,35	0,66	0,57
0,3	0,7	0,45	0,31	0,66	0,44
0,4	0,5	0,45	0,34	0,66	0,60
0,4	0,6	0,45	0,34	0,66	0,60
0,4	0,7	0,45	0,31	0,66	0,44
0,5	0,5	0,43	0,27	0,61	0,49
0,5	0,6	0,43	0,27	0,61	0,49
0,5	0,7	0,43	0,26	0,61	0,45
0,6	0,6	0,43	0,27	0,61	0,49
0,6	0,7	0,43	0,26	0,61	0,45