КЛАССИФИКАЦИЯ ВАКАНСИЙ С ЦЕЛЬЮ ПОСЛЕДУЮЩЕЙ ОПТИМИЗАЦИИ ПУБЛИКАЦИИ ОБЪЯВЛЕНИЙ

М. В. Быстрова, И. А. Адуцкевич

В эпоху современной глобализации ситуация на рынке труда такова, что компании желают видеть на рабочих местах сотрудников с опытом работы и определенным набором знаний. Но такого рода специалисты, как правило, уже трудоустроены в других компаниях и ищут работодателя, способного предложить более выгодные условия труда.

По статистике, заполнение большинства рабочих мест происходит с помощью публикаций объявлений о вакансиях. Успех такого набора зависит от того, как компания преуспела в составлении соответствующего объявления. Важно знать критерии, которые способны заинтересовать потенциального сотрудника, или же, другими словами, ценность предложения для работника. Данные критерии можно выделить, проанализировав все вакансии, на которые идет отклик.

При реализации алгоритма определения порядка близости между публикациями вакансий была использована технология Doc2Vec, предназначенная для обработки текстовых данных и представления их в виде векторного пространства.

ТЕХНОЛОГИЯ DOC2VEC

Данная технология собирает статистику совместного появления слов в фразах, а после этого при помощи нейронных сетей решает задачу снижения размерности, формируя на выходе компактные векторные представления слов максимально учитывая степень отношения этих слов в обрабатываемом тексте. [2].

В Doc2Vec используется нейронная сеть прямого распространения, на вход которой должны подаваться векторы фиксированной длины. Для того, чтобы было возможно применять данную технологию для работы с текстами разной длины, выполняется приведение векторов к одной размерности. При таком усреднении учитывается порядок слов благодаря добавлению вектора-абзаца и вектора-документа.

Технология Doc2Vec базируется на двух методах (см. Рисунок 1):

- Distributed Memory/распределенная память (DM) – прогнозирует слово по известным предшествующим словам и вектору абзаца;
- Distributed Bag of Words/распределенный мешок слов (DBOW) – прогнозирует случайные группы слов в абзаце на основании вектора абзаца.

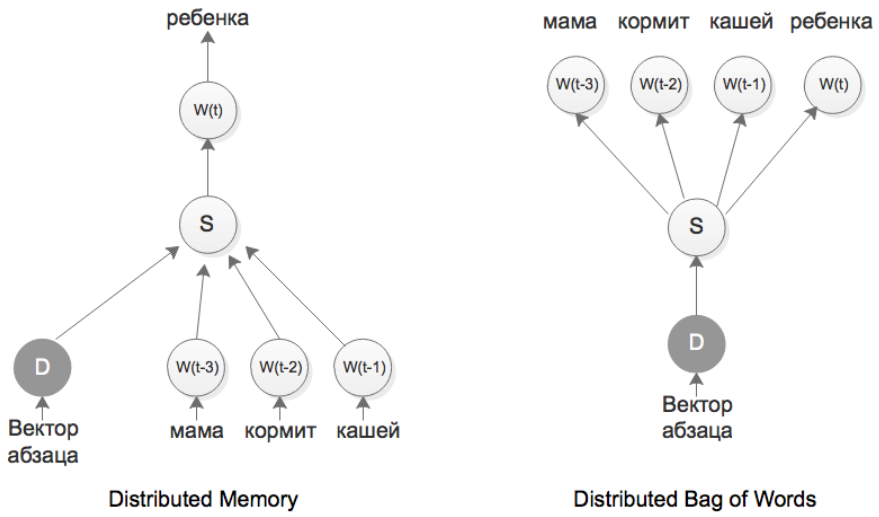

Рис.1. Архитектура методов Distributed Memory и Distributed Bag of Words

Данная технология хорошо подходит для анализа небольших документов (таких как рецензии или же публикации вакансий), в которых необходимо анализировать слова не по отдельности, а в рамках всего документа.

ОСНОВНЫЕ ЭТАПЫ ПОСТРОЕНИЯ КЛАССИФИЦИРУЮЩЕЙ МОДЕЛИ

ЭТАП 1. Предварительная обработка и индексация

Под предварительной обработкой понимают процесс преобразования последовательности слов, встречающихся в документе, в n-мерное пространство [1]. При этом происходит извлечение наиболее значимых слов (термов), удаление стоп-слов и html-тегов, а также и более сложные формы обработки текстовой информации такие, как морфологический и синтаксический анализ документов. Следует отметить, что многие классифицирующие алгоритмы требуют подачи на вход предварительно размеченной выборки.

ЭТАП 2. Построение и обучение классификатора

После формирования и предварительной обработки тренировочного набора документов следует выбор и построение классифицирующей модели. На вход данной модели подается тренировочная выборка для формирования словаря и дальнейшего обучения. Качество классифицирующей модели зависит от условий поставленной задачи и сформированной тренировочной выборки.

ЭТАП 3. Оценка качества работы классификатора

Для оценки качества работы классификатора используется тестовая выборка документов, для которых заранее определены классы. Ее подают на вход классификатора, после чего используется одна из следующих характеристик: *доля правильных ответов (accuracy)*, *полнота (recall)*, *точность (precision)* и *F-мера*.

Следует отметить, что эффективность работы классификатора напрямую зависит от размера и качества предварительной обработки данных, выступающих в качестве тренировочной выборки, а также метода построения самого классификатора.

АЛГОРИТМ ОПРЕДЕЛЕНИЯ ПОРЯДКА БЛИЗОСТИ МЕЖДУ ПУБЛИКАЦИЯМИ ВАКАНСИЙ НА БАЗЕ ТЕХНОЛОГИИ DOC2VEC

Для моделирования векторного пространства, языка программирования Python, разработанный в рамках данной работы алгоритм представляет собой веб-сервис, работающий по протоколу HTTP. Сама модель построена с использованием набора инструментов Gensim.

Имеется некоторая выборка, состоящая из 10 тысяч документов, разбитых более, чем на 20 классов (Accounting, Audit, Design, HR, Nursing и т.д.). Каждый документ проходит через этапы предварительной обработки: удаляются html теги и знаки пунктуации. Далее каждый документ преобразуется в массив слов, который размечается согласно требованиям алгоритма Doc2Vec. При этом, помечается принадлежность документа определенному классу. Например, для класса «Audit» – tags=['AUDIT_#документа'].

После создания модели, формирования словаря и обучения, в модели хранятся векторы документов с наиболее существенными словами, которые называются термами. В рамках реализованной модели, каждому документу соответствует ровно сто термов с весовыми коэффициентами, определяющими значимость в рамках отдельного документа. Также можно выделить термы, соответствующие определенному классу документов.

Например, для класса «Audit» получили набор термов с весовыми коэффициентами:

[(u'financial', 0.6264161467552185), (u'english', 0.5597227215766907), (u'management', 0.5499826669692993), (u'communication', 0.5458555221557617), (u'presentation', 0.5436646938323975), (u'excel', 0.5335297584533691), (u'analyse', 0.5295985341072083), (u'engineer', 0.5220131874084473), (u'experience', 0.5190562009811401), (u'sql', 0.5189145803451538)]

Представленные выше термы, в той или иной степени отражены в большинстве документов класса. Их также можно получить и проанализировать.

Можно также определить, какие документы являются похожими в рамках данной модели, а также найти похожие термы по всем документам.

Например, для документа, помеченного *AUDIT_1* результаты поиска близких по смыслу ранее опубликованных вакансий:

[('AUDIT_7', 0.4295151162147515), ('AUDIT_19', 0.3395153326147522), ('AUDIT _3', 0.314151162736152), ('AUDIT _28', 0.304158459836152), ('AUDIT_33', 0.297155656064251), ('AUDIT _10', 0.296252651984625), ('AUDIT _60', 0.2844726321794921), ('AUDIT_98', 0.2815384230262146), ('AUDIT_5', 0.2799204210281372), ('AUDIT_87', 0.2797151162147542)]

Видно, что поиск схожих публикаций вакансий в рамках класса, которому принадлежит документ, т. к. в рамках модели они считаются наиболее схожими.

ЗАКЛЮЧЕНИЕ

Как было ранее отмечено, анализ публикации вакансий имеет большую практическую значимость для анализа рынка вакансий, так и для компаний, заинтересованных в приеме на работу квалифицированных специалистов. С задачей анализа публикаций хорошо справился алгоритм на базе технологии Doc2Vec, который позволил выявить набор наиболее значимых термов, как в рамках отдельно взятой публикации, так и в рамках определенного класса вакансий. Отличительной особенностью использованной технологии является то, что она работает с заранее размеченным набором документов, что позволяет обрабатывать документ в рамках определенного класса, а не всей тренировочной выборки.

Планируется также классифицирование резюме и, в дальнейшем, разработка алгоритма с целью определения соответствия резюме предоставляемой вакансии и наоборот.

Литература

1. *Агеев, М. С.* Методы автоматической рубрикации текстов, основанные на машинном обучении знаниях экспертов / М. С. Агеев. - М.: Либроком (Editorial URSS), 2004. – 106 с.
2. *Tomas Mikolov.* Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS / Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, 2013.