Анализ неструктурированных данных

Домашняя работа 2

Идентификация парафраза

Дедлайн: 7.11.2018

Результатом выполнения задания является ipython-ноутбук, в котором представлены все скрипты, реализующие проделанные эксперименты и отчет, описывающий каждый шаг и всю логику вычислений и все полученные результаты. Отчёт также должен содержать краткое описание всех использованных моделей (с указанием на цитируемый источник описания модели, если такой имеется) и инструментов. Вся проделанная работа должна быть понятна из этого текста.

На усмотрение проверяющего остается штраф (т.е. снижение оценки) за неаккуратное оформление, копирование Википедии и любого другого ресурса без указания источника, списывание, неформальный стиль изложения и обилие стилистических ошибок.

Домашнее задание выполняется в группах по 1-3 человека.

Все вопросы по содержанию, оформлению и сдаче домашнего задания можно задавать учебным ассистентам в чате АНД 2018.

Выполненное домашнее задание сдается через систему AnyTask. Инструкции по использованию системы AnyTask будут дополнительно опубликованы в телеграм-канале.

Домашнее задание посвящено идентификации парафраза. Под парафразом мы понимаем близость по смыслу двух предложений. Например, мы можем сказать, что парафразом являются два новостных заголовка “французы стали чемпионами мира” и “Итоги мундиаля: Франция встречает чемпионов”.

Мы можем сформулировать несколько подходов к определению парафраза. Пусть дано некоторое множество предложений $\mathbb{S}$.

1. unsupervised: для любых $s_1, s_2 \in \mathbb{S}$ по некоторому принципу задаем функцию близости $\texttt{sim}(s_1, s_2)$ и устанавливаем порог на ее значения. Считаем, что пара предложений является парафразом, если значения функции близости выше заданного порога.

2. supervised: для каждой пары $s_1, s_2 \in \mathbb{S}$ дана метка $l \in [0, 1]$. Тогда задача формулируется как задача классификации. В свою очередь, к решению этой задачи можно подойти одним из многих способов. Широко используются два подхода:

   - обучение независимых от задачи эмбеддингов предложений и использование методов классификации (логистическая регрессия, SVM и др.)
   - обучение нейронной сети и эмбеддингов предложений одновременно. На вход нейронной сети в этом случае подают пары предложений в виде эмбеддингов слов и метки.

В этой работе вам предстоит сравнить между собой все три подхода к решению задачи парафраза.

Данные для задания: русскоязычный корпус парафраза ParaPhraser (`http://paraphraser.ru/download/get?file_id=1`).

Разбейте корпус случайным образом на тестовую и обучающую выборки в отношении 3:1.

Сравните все реализованные методы по accuracy и f-мере.

Задание 1 (4 балла) Unsupervised метод

В качестве базового метода измерения расстояния, используйте Word Mover's Distance [1]. Для этого метода вам понадобится модель эмбеддингов, поэтому вы можете использовать любую известную вам предобученную модель или обучить самостоятельно новую. Фактически, в этой части задания вам нужно эмпирически подобрать значения порога на близость.

Задание 2 (2 балла + количество моделей эмбедингов / 2) Supervised I: бинарная классификация

Постройте несколько моделей эмбеддингов предложений, например, усредненный w2v, усредненный w2v с весами, doc2vec, ELMo [2, 3]. Используйте любой метод бинарной классификации и сравните, какая модель эмбеддингов доставляет большее значение accuracy и f-меры при прочих равных.

Задание 2 (4 балла) Supervised II: Siamese DAN

Реализуйте сиамскую Deep Averaging Network [4]: у сети два входа, на каждый подается усредненный w2v для одного из двух предложений. Каждая из двух частей нейронной сети имеет несколько полносвязных скрытых слоев. Выходы обеих частей подаются на вход итоговым нескольких полносвязным слоям. Выход сети решает бинарную задачу классификации.

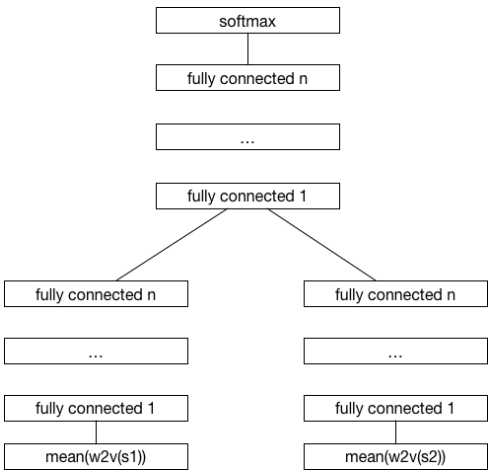

Рис. 1: Siamese Deep Averaging Network для идентификации парафраза

Задание 4 (бонус, до 3 баллов) Визуализация

Придумайте способ красиво и информативно визуализировать решение задачи.

Ссылки

1. Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. "From word embeddings to document distances."In International Conference on Machine Learning, pp. 957-966. 2015.

2. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

3. Обученная модель: `http://docs.deeppavlov.ai/en/master/intro/pretrained_vectors.html`

4. Iyyer, M., Manjunatha, V., Boyd-Graber, J. and Daumé III, H., 2015. Deep unordered composition rivals syntactic methods for text classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Vol. 1, pp. 1681-1691).