Национальный исследовательский университет – «Высшая школа экономики»
Аспирантская школа по компьютерным наукам.

Рецензия на доклад Анны Потапенко, аспиранта второго года (Департамент больших данных и информационного поиска),
на тему:
«Вероятностное тематическое моделирование и нейросетевые модели языка для обучения векторного представления слов, контекстов и документов»

Докладчик рассказал о современных подходах обучения моделей векторного представления текстовой информации: слов, контекстов и документов. Были рассмотрены два направления: вероятностное тематическое моделирование и нейросетевые модели языка. К первому направлению относятся LDA, PLSA, VSMS и другие. Например, для VSMS основными моментами являются использование дистрибутивной гипотезы (слова, использованные в одних и тех же контекста, подразумевают вероятнее всего один и тот же смысл), обучение без учителя, определение близости между словами, фразами, предложениями и даже целыми абзацами. Все эти методы основаны на предсказании вероятности N-грам по выбранному корпусу. Ко второму направлению относятся Word embeddings, Paragraph2vec, CBOW и другие. Все они обычно используют двуслойную нейронную сеть, которую обучают на корпусе.

Для каждого метода даны базовые определения, впоследствии докладчик привел преимущества и недостатки каждого из них. Слушателям был представлен пример, где на практике определялись имя/пол автора текста. В конце доклада Анна Потапенко рассказала о достижениях в работе за прошедший год и планах на следующий год.

В докладе была рассмотрена очень актуальная тема, приведены самые свежие работы и публикации. Хотелось бы увидеть в будущем какое-либо практическое применение описанных методов. Единственное замечание к докладчику – отсутствие списка литературы, из которого можно подчерпнуть больше информации.

Кирилл Малахов

01.03.2016