Московский государственный технический университет им. Н.Э. Баумана

Кафедра «Системы обработки информации и управления»
ИУ-5

Методы анализа данных

Гапанюк Юрий Евгеньевич, к.т.н., доцент кафедры ИУ-5

ЛЕКЦИЯ №2

Векторизованное представление текстов

План

1. Векторизованное представление.
2. Векторизованное представление текстов.
3. Примеры решения практических задач на основе векторизованного представления текстов.
4. Семинар. Векторизованное представление текстов с использованием библиотек на языке Python.

Векторизованное представление

- Операция преобразования (кодирования) сведений о каком-либо объекте в многомерное векторное пространство называется «векторным представлением» этого объекта.
- В англоязычной литературе используется термин «embedding».
- Мы также будем использовать термин «эмбеддинг».
- Эмбеддинг осуществим для объектов произвольного вида – табличных данных, текстов, изображений, графов и т.д.
- Совместное кодирование данных об одном и том же объекте в различных представлениях (текст, изображение, …) называют «мультимодальным эмбеддингом». Он является основной мультимодального обучения.

Векторизованное представление текстов

- Можно разделить на два периода:
  1. До эпохи машинного обучения.
  2. В эпоху машинного обучения.
- До эпохи машинного обучения в основном использовался подход на основе кодирования текстов экспертами.
  - В. А. Тузов - "КОМПЬЮТЕРНАЯ СЕМАНТИКА РУССКОГО ЯЗЫКА“
  - В.В. Мартынов - Универсальный семантический код (УСК)
- В эпоху машинного обучения теоретической базой для векторных представлений является дистрибутивная семантика. Векторные представления формируются автоматически на основе методов машинного обучения.

Векторизованное представление текстов

- Векторная модель слов и n-грамм (разбирается на семинаре). Данная модель не учитывает семантику слов.
- Модели Word2Vec и Glove учитывают семантику. Модель fastText также учитывает фрагменты слов.
- Эмбеддинги на основе сложных нейросетевых моделей. Наиболее активно применяется BERT (статьи с пояснением работы 1 и 2).
- deeppavlov.ai – пример диалоговой системы на основе векторизованного представления.

Задача кластеризации текстовых новостей

1. Алгоритм кластеризации новостного потока текстовых сообщений. (препринт)

2. Улучшения алгоритма кластеризации новостного потока текстовых сообщений. (препринт)

Задача анализа тональности текста

- Классическая задача анализа тональности текста.

- Аспектно-ориентированный анализ тональности текста на естественном языке. (препринт)

Задача разработки гибридной диалоговой системы

- Гибридная интеллектуальная русскоязычная диалоговая информационная система на основе метаграфового подхода. (препринт)