АННОТАЦИЯ К РАБОЧЕЙ ПРОГРАММЕ ДИСЦИПЛИНЫ
«Методы моделирования естественных языков»

по основной профессиональной образовательной программе по направлению подготовки 09.06.01 «Информатика и вычислительная техника» (уровень подготовки научно-педагогических кадров)

Направленность (профиль): Математическое моделирование, численные методы и комплексы программ
Общий объем дисциплины – 4 з.е. (144 часов)
Форма промежуточной аттестации – Зачет.
В результате освоения дисциплины обучающийся должен обладать следующими компетенциями:

- ОПК-1: Владением методологией теоретических и экспериментальных исследований в области профессиональной деятельности;
- ОПК-2: Владением культурой научного исследования, в том числе с использованием современных информационно-коммуникационных технологий;
- ПК-2: способность проводить комплексные исследования научных и технических проблем с применением современных технологий математического моделирования и вычислительного эксперимента;

Содержание дисциплины:
Дисциплина «Методы моделирования естественных языков» включает в себя следующие разделы:
Форма обучения очная. Семестр 6.
1. Введение в модели и алгоримы обработки естественных языков. Природа сложности задач автоматической обработки текстов (АОТ)
на естественном языке. Модели представления знаний.
Проблемы контекстной зависимости и многозначности.
Синтаксис и семантика естественного языка. Общие требования к семантической модели.
Статистические модели. Векторные представления.
Модели машинного обучения. Сверточные и рекуррентные нейросети. Архитектура LSTM..
2. Фаза предварительной обработки текста. Токенизацию. Удаление стоп-слов.
Приведение слов к единому регистру. Устранение шума.
Лемматизация. Обработка аббревиатур, сленга и коррекция ошибок..
3. Тезаурусы и лексиконы. Тезаурус WordNet. Классическая модель WordNet.
Семантические отношения. Синсеты. Русскоязычные тезаурусы РуТез и RuWordNet.
Лексикон. Тональный лексикон..
4. Статистические модели. Дистрибутивная гипотеза и понятие о близости значений слов.
Частотность. Взвешивание. Вес в пределах коллекции.
Векторные модели и плотные вектора в Word Embeddings.
Модели индексации: TF, TF-IDF, GloVe, Word2Vec.
Слово и его окрестность. Реализации word embeddings на основе
метода GloVe, модели Skip-gram и CBOW (continuous
Bag-of-words) в составе алгоритма Word2Vec.
Подходы к определению окрестности слова в модели:
предсказание слова по его окрестности в CBOW,
предсказание окрестности по слову в Skip-Gram..
5. Синтаксический и семантический анализ. Понятие грамматики. Контекстная зависимость. Контекстно-свободные грамматики.
Понятие дерева вывода. Естественные и формальные языки.
Семантический парсер естественного языка RML.
Ресурсы, применяемые в задачах АОТ.
Язык Питон и библиотеки обработки естемтаенных языков..
6. Модели в задаяах аннотирования текстов. Общее и тематически-ориентированное аннотирование.

Извлекающие и генерирующие модели автоматического аннотирования.
Генерирующие методы на основе логико-семантических отношений между фрагментами текста.
Извлекающие методы на основе лексических цепочек, латентно-семантическом анализе, скрытом распределении Дирихле..

7. Модели в задачах классификации текстов. Задача классификации. Понятие аспекта и категории.
Предобработка и индексация, уменьшение размерности пространства признаков, построение и обучение классификатора, оценка качества классификации.
Классификатор на основе метода логистической регрессии, наивного байесовского классификатора (NBC), k-ближайших соседей (KNN), метода опорных векторов (SVM), деревьев решений и случайных лесов..

Разработал:
профессор
кафедры ПМ Е.Н. Крючкова
Проверил:
Декан ФИТ А.С. Авдеев