Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет

«Высшая школа экономики»

Факультет гуманитарных наук

Школа лингвистики

Программа дисциплины

Машинное обучение

для направления 45.04.03 «**Фундаментальная и прикладная лингвистика»** подготовки **магистра**

Авторы программы:

Черняк Е.Л., преподаватель (echernyak@hse.ru)

Одобрена на заседании Школы лингвистики ФГН «5» июня 2017 г., протокол №11
Руководитель Школы лингвистики Е.В. Рахилина _____ [подпись]

Рекомендована Академическим советом образовательной программы
«22» июня 2017 г., Протокол № 10

Утверждена «22» июня 2017 г.

Академический руководитель образовательной программы
Бонч-Осмоловская А. А. _____ [подпись]

Москва, 2017

1. Аннотация

Дисциплина «Машинное обучение» предназначена для подготовки магистров направления 45.04.03 «Фундаментальная и прикладная лингвистика». Она продолжает цикл дисциплин, связанных с основами анализа данных, информационных технологий и программирования.

В курсе изучаются основные методы машинного обучения, используемые для решения задач автоматической обработки текстов. Затрагиваются задачи классификации текстов на два и более классов, многотемной и иерархической классификации. Определяется класс задач АОТ, решаемых с помощью методов классификации последовательностей. Рассматриваются методы снижения размерности, основанные как на матричных разложениях, так и на обучении векторам слов. Определяются цели и основные методы кластерного анализа. Представляется введение в современные методы глубинного обучения.

Теоретический материал курса подкрепляется практическими занятиями по использованию популярных инструментов по изучаемой тематике.

2. Область применения и нормативные ссылки

Настоящая программа устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов первого года обучения в магистратуре по направлению «Фундаментальная и прикладная лингвистика». Дисциплина является курсом по выбору.

Программа разработана в соответствии с:

• Образовательным стандартом НИУ ВШЭ;

• Образовательной программой подготовки магистра по направлению 45.04.03 «Фундаментальная и прикладная лингвистика».;

• Рабочим учебным планом подготовки магистра по направлению 45.04.03 «Фундаментальная и прикладная лингвистика»., утвержденным в 2016 г.

3. Цели освоения дисциплины

Данная дисциплина ставит своей целью изучение основных задач и методов машинного обучения, а также их использования для решения задач автоматической обработки текстов, наряду с освоением программных систем и инструментов, в которых реализованы данные методы. Эти базовые знания и навыки необходимы в профессиональной деятельности специалистов по компьютерной лингвистике.

4. Компетенции, формируемые в результате освоения дисциплины

В результате изучения дисциплины студенты должны:

• Знать основные задачи и методы машинного обучения с учителем и без;

• Уметь самостоятельно формулировать задачи классификации текстов или предложений и уметь выбирать подходящий алгоритм классификации, а так же пользоваться его готовыми реализациями;

• Уметь самостоятельно выбирать подходящий метод снижения размерности;

• Уметь самостоятельно формулировать задачу кластеризации текстов или предложений и уметь выбирать подходящий алгоритм классификации, а так же пользоваться его готовыми реализациями;

• Знать основные классы и методы библиотек scikit-learn и NLTK.

В результате изучения дисциплины студент осваивает и развивает следующие компетенции:

Компетенция	**Код по ФГОС/ НИУ**	**Дескрипторы – основные признаки освоения (показатели достижения результата)**	**Формы и методы обучения, способствующие формированию и развитию компетенции**
Умение работать на компьютере, навыки использования основных классов программного обеспечения, работы в компьютерных сетях	ИК-2	Студент демонстрирует владение интерфейсом программных систем для обработки и анализа текстов	Выполнение домашних заданий, ориентированных на использование программных систем обработки и анализа текстов
Способность решать задачи производственной и технологич. деятельности на профессион. уровне, включая разработку математических моделей, алгоритмических и программных решений	ПК-8	Студент демонстрирует компетентность в выборе той или иной программной системы для решения поставленной перед ним задачи обработки и анализа текстов	Лекции по основным задачам и методам обработки и анализа текстов; решение задач, требующих выбор метода обработки и анализа текстов и программной системы, в которой данный метод реализован
Способность применять в профессиональной деятельности современные языки программирования и языки баз данных, операционные системы, электронные библиотеки и пакеты программ и т.п.	ПК-9	Студент демонстрирует понимание основных методов обработки и анализа текстов, владение основными программными системами обработки и анализа текстов	Лекции по основным задачам и методам обработки и анализа текстов; домашние задания, ориентированные на использование программных систем обработки и анализа текстов

5. Место дисциплины в структуре образовательной программы

Настоящая учебная дисциплина является курсом по выбору и входит в цикл дисциплин информационных технологий в учебной программе подготовки магистра направления 45.04.04 «Фундаментальная и прикладная лингвистика».

Изучение курса «Машинное обучение» требует базовых знаний по математике (в объеме адаптационных курсов магистерской программы первого года обучения по направлению). Необходимо также владение базовыми навыками программирования на языке высокого уровня (в объеме обязательного курса «Программирование (язык Python)» первого года обучения указанной магистерской программы).

Основные положения дисциплины «Машинное обучение» должны быть использованы при выполнении проектных, курсовых и выпускных квалификационных работ.

6. Тематический план дисциплины «Программные системы обработки и анализа текстов»

№	Название темы	Всего часов по дисциплине	Аудиторные часы		Самостоятельная работа
			Лекции	Сем. и практика занятия	
1	Введение		1	2	2
2	Задача бинарной классификации. Машины опорных векторов. Фильтрация спама.		1	2	8
3	Задача классификации на несколько классов. Метод наивного Байеса, деревья решений. Тематическая классификация и классификация по тональности.		1	2	12
4	Многотемная классификация, иерархическая классификация. Тематическая классификация – продолжение.		1	4	12

5	Классификация последовательностей. Морфологический анализ и разрешение омонимии. Выделение именованных сущностей.		2	2	10
6	Методы снижения размерности. Вероятностное тематическое моделирование. Вектора слов.		2	2	10
7	Введение в глубинное обучение. Рекуррентная нейронная сеть. Сверточные нейронные сети. Архитектуры Senna и SyntaxNet		2	4	14
8	Кластерный анализ. Кластеризация новостного потока.		2	2	12
	Итого	114	12	20	82

7. Формы контроля знаний студентов

Курс «Программные системы обработки и анализа текстов» читается в 3 модуле.

Тип контроля	Форма контроля	Параметры
Текущий контроль	Домашние задания (3)	
Итоговый контроль в 3 модуле	Устный экзамен	120 минут, задаются вопросы по билетам

Критерии оценки знаний

На итоговом контроле студент должен продемонстрировать владение основными понятиями из пройденных тем дисциплины.

Итоговый контроль проводится в форме устного экзамена, включающего несколько вопросов по темам дисциплины.

Порядок формирования оценок по дисциплине

Преподаватель оценивает самостоятельную работу студентов по выполнению домашних работ, выдаваемых на практических занятиях – при этом оценивается правильность выбора метода решения задачи и эффективность его использования. Оценки за домашние задания выставляются в рабочую ведомость, и перед экзаменом модуля за домашние задания выставляется результирующая оценка по десятибалльной шкале $O_{сам.\ работа}$.

Оценка итогового контроля выставляется по следующей формуле:

$$O_{диицциплина} = 0{,}5 \cdot O_{экзамен} + 0{,}5 \cdot O_{сам.\ работа}$$

и округляется до целого числа арифметическим способом,

где $O_{экзамнг}$ – оценка за работу непосредственно на устном экзамене.

В случае пропусков занятий и домашних заданий студент может сдать все домашние задания не позднее чем за 5 дней до экзамена – в этом случае они учитываются описанным выше способом.

В диплом выставляется ***результирующая оценка*** $O_{дисциплина}$ по данной учебной дисциплине.

8. Содержание программы по темам

Тема 1. Введение

1. Связь машинного обучения и автоматической обработки текстов.

2. Обзор существующих задач автоматической обработки текстов и соответствующих им методов машинного обучения.

3. Обзор существующих систем обработки и анализа текстов. Классификация систем обработки и анализа текстов.

Основная литература

1. Machine learning methods for natural language processing. Michael Collins. // http://www.cs.columbia.edu/~mcollins/papers/tutorial_colt.pdf (онлайн ресурс)

2. Christopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

Дополнительная литература

1. Sparck Jones, K. Natural language processing: a historical review // Current Issues in Computational Linguistics: in Honour of Don Walker. – 1994. – С. 3-16

Тема 2. Задача бинарной классификации. Машины опорных векторов. Фильтрация спама.

1. Формальная постановка задачи классификации на два класса.

2. Векторная модель представления текста. Пространство слов. Простейшая классификация: метод ближайшего соседа. Меры качества классификации.

3. Понятие о разделяющей гиперплоскости и опорных векторах. Ядерные функции. Машины опорных векторов.

4. Практика: фильтрация спама. Библиотека scikit-learn.

Основная литература

1. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

2. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. — Springer, 2009

Дополнительная литература

1. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

2. Thiago S. Guzella, Walmir M. Caminhas A review of machine learning approaches to Spam filtering // Expert Systems with Applications. — 2009. — Vol. 36, no. 7.

Тема 3. Задача классификации на несколько классов. Метод наивного Байеса, деревья решений. Тематическая классификация и классификация по тональности.

1. Формальная постановка задачи классификации на несколько классов. Примеры классификации на несколько классов: классификация по теме, классификация по тональности.

2. Байесовский подход к классификации. Априорные и апостериорные вероятности. Преобразование Лапласса.

3. Деревья решений. Коэффициенты связи целевого признака с объясняющими качественными и количественными признаками. Принцип минимизации энтропии. Алгоритм ID3. Проблема переобучения.

4. Практика: тематическая классификация или классификация по тональности. Библиотека scikit-learn.

Основная литература

1. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

2. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. — Springer, 2009

Дополнительная литература

1. Frank, Eibe, and Remco R. Bouckaert. "Naive bayes for text classification with unbalanced classes." European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg, 2006.

2. Aggarwal, Charu C., and ChengXiang Zhai. Mining text data. Springer Science & Business Media, 2012.

3. Korde, Vandana, and C. Namrata Mahender. "Text classification and classifiers: A survey." International Journal of Artificial Intelligence & Applications 3.2 (2012): 85.

Тема 4. Многотемная классификация, иерархическая классификация. Тематическая классификация – продолжение.

1. Примеры задач многотемной классификации и иерархической классификации. Связь с классификацией на несколько классов.

2. Практика: тематическая классификация – продолжение.

Основная литература

1. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

2. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. — Springer, 2009

Дополнительная литература

1. Tsoumakas, Grigorios, and Ioannis Katakis. "Multi-label classification: An overview." Dept. of Informatics, Aristotle University of Thessaloniki, Greece (2006).

2. Sun, Aixin, and Ee-Peng Lim. "Hierarchical text classification and evaluation." Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001.

Тема 5. Классификация последовательностей. Морфологический анализ и разрешение омонимии. Выделение именованных сущностей.

1. Формальная постановка задачи классификации последовательностей. Примеры классификации последовательностей: морфологическая разметка, разрешение морфологической неоднозначности, извлечение именованных сущностей.

2. Марковские цепи. Скрытые цепи Маркова. Алгоритм Витерби.

3. Графовые модели. Условные случайные поля.

4. Практика: выделение именованных сущностей. Библиотекса CRFSuite.

Основная литература

1. Xing, Zhengzheng, Jian Pei, and Eamonn Keogh. "A brief survey on sequence classification." ACM SIGKDD Explorations Newsletter 12.1 (2010): 40-48.

2. Zhou, GuoDong, and Jian Su. "Named entity recognition using an HMM-based chunk tagger." proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002.

3. Sha, Fei, and Fernando Pereira. "Shallow parsing with conditional random fields." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

Дополнительная литература

1. McCallum A. Efficiently inducing features of conditional random fields. In Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence 2002 Aug 7 (pp. 403-410). Morgan Kaufmann Publishers Inc.

2. Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26.

3.

Тема 6. Методы снижения размерности. Вероятностное тематическое моделирование. Вектора слов.

1. Необходимость снижения размерности. Сингулярное разложение матрицы.

2. Методы выделения скрытых тем: латентный семантический анализ, латентное разложение Дирихле. Параметры модели. Выбор числа скрытых тем. Библиотека Gensim.

3. Вектора слов. Связь между между методами вычисления векторов слов и методами снижения размерности. Перцептрон. Метод обратного распространения ошибки. Дистрибутивная гипотеза. Вероятностная языковая модель. Word2vec, архитектуры skip-gram и CBOW. Библиотека Gensim.

Основная литература

1. Воронцов, К. В. "Вероятностное тематическое моделирование." Москва (2013).

2. Mikolov, T., and J. Dean. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems (2013).

Дополнительная литература

1. Deerwester, Scott, et al. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391.

2. Hofmann, Thomas. "Probabilistic latent semantic indexing." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.

3. Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. "Multimodal Distributional Semantics." J. Artif. Intell. Res.(JAIR) 49.1-47 (2014).

4. Bengio, Yoshua, et al. "A neural probabilistic language model." journal of machine learning research 3.Feb (2003): 1137-1155.

5. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." EMNLP. Vol. 14. 2014.

6. Levy, Omer, and Yoav Goldberg. "Dependency-Based Word Embeddings." ACL (2). 2014.

Тема 7. Введение в глубинное обучение. Сверточные нейронные сети. Выделение фактов и событий. Архитектуры Senna и SyntaxNet

1. Модель однослойной нейронной сети

2. Связь вероятностных языковых моделей и дистрибутивной семантики.

3. Непрерывные вероятностные языковые модели. Многоуровневные нейронные сети.

4. Сверточные нейронные сети и рекуррентные нейронные сети. Их использование в задачах автоматической обработки текстов. Практика: извлечение фактов с помощью сверточной нейронной сети.

5. Архиктура Senna. Практика: извлечение именованных сущностей.

6. Архиктура SyntaxNet. Практика: синтаксический анализ.

Основная литература

1. Ponte, Jay M., and W. Bruce Croft // A language modeling approach to information retrieval. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998.

2. Bengio, Yoshua. Learning deep architectures for AI. // Foundations and trends® in Machine Learning 2.1 (2009): 1-127.

3. Collobert, Ronan, et al. Natural language processing (almost) from scratch. // Journal of Machine Learning Research 12.Aug (2011): 2493-2537.

4. Mikolov, Tomas, et al. Efficient estimation of word representations in vector space. // arXiv preprint arXiv:1301.3781 (2013).

Дополнительная литература

1. Bengio, Yoshua, et al. A neural probabilistic language model // Journal of machine learning research 3.Feb (2003): 1137-1155.

2. Bruni, Elia, Nam-Khanh Tran, and Marco Baroni // Multimodal Distributional Semantics. J. Artif. Intell. Res.(JAIR) 49.1-47 (2014).

3. Levy, Omer, and Yoav Goldberg. Dependency-Based Word Embeddings. // ACL (2). 2014.

4. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. Glove: Global Vectors for Word Representation. // EMNLP. Vol. 14. 2014.

Тема 8. Кластерный анализ

1. Задача кластерного анализа. Признаковое описание текста. Вычисление близости между текстами.

2. Метрическая классификация: метод k-means. Агломеративная иерархическая классификация. Спектральная классификация: метод normalized cut.

3. Практика: классификация новостного потока.

Основная литература

1. Mirkin, Boris. Core concepts in data analysis: summarization, correlation and visualization. Springer Science & Business Media, 2011.

Дополнительная литература

2. Moisl, Hermann. Cluster analysis for corpus linguistics. Vol. 66. Walter de Gruyter GmbH & Co KG, 2015.

3. McKeown, Kathleen R., et al. "Tracking and summarizing news on a daily basis with Columbia's Newsblaster." Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., 2002.

9. Образовательные технологии

В преподавании данной дисциплины сочетаются:

• лекции в традиционной форме;

• практические занятия, в ходе которых студенты осваивают основные программные системы обработки и анализа текстов;

• домашние практические задания по использованию программных систем машинного обучения обработки и анализа текстов по всем основным темам дисциплины.

10. Оценочные средства для текущего и итогового контроля

Примеры домашних работ

1. Сравнить эффективность использования различных ядерных функций для машины опорных векторов в задаче фильтрации спама.

2. Использовать метод наивного Байеса для классификации коллекции Reuters. Оценить связь качества классификации с выбираемыми параметрами.

3. Использовать деревья решений для классификации коллекции Reuters. Оценить связь качества классификации с выбираемыми параметрами.

4. Реализовать функцию вычисления похожести слов, используя библиотеку gensim и различные меры близости и методы снижения размерности.

5. Предположить, что в собственной коллекции текстов на русском языке существует некоторое число скрытых тем. Выделить их, используя метод LDA. Объяснить полученные результаты.

6. Разметить текст на русском с помощью синтаксического парсера SyntaxNet. Построить деревья зависимостей.

7. Из данного текста выделить все именованные сущности, используя библиотеку Senna и ее интерфейс в NLTK.

Вопросы для оценки качества освоения дисциплины

Тема 1.

1. Перечислите основные методы машинного обучения.

2. Что такое обучающее, тестовое и отладочное множество?

3. Приведите пример задачи автоматической обработки текстов, опишите примерную схему ее решения с помощью машинного обучения.

4. Назовите несколько библиотек для машинного обучения.

5. Назовите несколько консольных приложений для машинного обучения.

Тема 2.

1. Объясните принципы векторной модели (VSM).

2. Что такое разделяющая гиперплоскость?

3. Что такое машина опорных векторов?

4. Какие ядерные функции вы знаете? Что такое kernel trick?

5. Приведите пример задачи бинарной классификации из области автоматической обработки текстов и из другой области.

Тема 3.

1. Перечислите основные принципы байесовского подхода к классификации.

2. Сколько алгоритмов построения деревьев решений вы знаете?

3. Существует ли геометрическая интерпретация методов деревьев решений?

4. Приведите пример задачи многомерной классификации из области автоматической обработки текстов и из другой области.

Тема 4.

1. Что такое многотемная классификация? Приведите пример задачи многотемной классификации из области автоматической обработки текстов и из другой области.

2. Что такое иерархическая классификация? Приведите пример задачи иерархическая классификации из области автоматической обработки текстов и из другой области.

Тема 5.

1. Какие методы классификации последовательностей вы знаете?

2. Что такое Марковская цепь? Как она связана с идей классификации последовательностей?

3. Что такое скрытая Марковская цепь? Как вычисляются параметры скрытой Марковской цепи?

4. Что такое условное случайное поле? Как вычисляются параметры условного случайного поля?

5. Что такое классификация последовательностей? Приведите пример задачи классификации последовательностей из области автоматической обработки текстов и из другой области.

Тема 6.

1. Почему возникает потребность в снижении размерностей?

2. Что такое сингулярное разложение? Можно ли использовать его для выделения скрытых тем?

3. Улучшает ли снижение размерности качество поиска?

4. Какие методы вероятностного моделирования вы знаете?

5. Какова связь между методами снижения размерности и вычислением векторов слов?

6. Какие виды векторов слов вы знаете?

Тема 7.

1. Требует ли обучение глубинной нейронной сложного отбора признаков?

2. Что такое нейронная сверточная сеть? В каких задачах целесобразно ее использование?

3. Что такое рекуррентная нейронная сеть? В каких задачах целесобразно ее использование?

4. В чем заключаются особенности архитектуры Senna? Для каких задач она используется?

5. В чем заключаются особенности архитектуры SyntaxNet? Для каких задач она используется?

Тема 8.

3. Дайте определение кластера.

4. Приведите пример задачи кластерного анализа из области автоматической обработки текстов и из другой области.

5. Что такое матрица расстояний (близости) и зачем она нужна? Какие способы вычисления расстояния (близости) между текстами вы знаете?

6. Перечислите основные шаги метода k-means.

7. Перечислите основные шаги метода normalized cut.

8.

Базовая литература

1. Steven Abney, Semisupervised Learning for Computational Linguistics. Chapman & Hall/CRC (Computer science and data analysis series, edited by David Madigan et al.), 2008

2. Daniel Jurafsky and James H. Martin. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Series in Articial Intelligence. Pearson Education International. Second Edition, 2009.

3. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

4. Christopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

Основная литература

1. Machine learning methods for natural language processing. Michael Collins. // http://www.cs.columbia.edu/~mcollins/papers/tutorial_colt.pdf (онлайн ресурс)

2. Christopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

3. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

4. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. — Springer, 2009

5. Xing, Zhengzheng, Jian Pei, and Eamonn Keogh. "A brief survey on sequence classification." ACM SIGKDD Explorations Newsletter 12.1 (2010): 40-48.

6. Zhou, GuoDong, and Jian Su. "Named entity recognition using an HMM-based chunk tagger." proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002.

7. Sha, Fei, and Fernando Pereira. "Shallow parsing with conditional random fields." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

8. Воронцов, К. В. "Вероятностное тематическое моделирование." Москва (2013).

9. Mikolov, T., and J. Dean. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems (2013).

10. Ponte, Jay M., and W. Bruce Croft // A language modeling approach to information retrieval. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998.

11. Bengio, Yoshua. Learning deep architectures for AI. // Foundations and trends® in Machine Learning 2.1 (2009): 1-127.

12. Collobert, Ronan, et al. Natural language processing (almost) from scratch. // Journal of Machine Learning Research 12.Aug (2011): 2493-2537.

13. Mikolov, Tomas, et al. Efficient estimation of word representations in vector space. // arXiv preprint arXiv:1301.3781 (2013).

14. Mirkin, Boris. Core concepts in data analysis: summarization, correlation and visualization. Springer Science & Business Media, 2011.

Дополнительная литература

1. Sparck Jones, K. Natural language processing: a historical review // Current Issues in Computational Linguistics: in Honour of Don Walker. – 1994. – С. 3-16

2. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

3. Thiago S. Guzella, Walmir M. Caminhas A review of machine learning approaches to Spam filtering // Expert Systems with Applications. — 2009. — Vol. 36, no. 7.

4. Frank, Eibe, and Remco R. Bouckaert. "Naive bayes for text classification with unbalanced classes." European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg, 2006.

5. Aggarwal, Charu C., and ChengXiang Zhai. Mining text data. Springer Science & Business Media, 2012.

6. Korde, Vandana, and C. Namrata Mahender. "Text classification and classifiers: A survey." International Journal of Artificial Intelligence & Applications 3.2 (2012): 85.

7. Tsoumakas, Grigorios, and Ioannis Katakis. "Multi-label classification: An overview." Dept. of Informatics, Aristotle University of Thessaloniki, Greece (2006).

8. Sun, Aixin, and Ee-Peng Lim. "Hierarchical text classification and evaluation." Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001.

9. McCallum A. Efficiently inducing features of conditional random fields. In Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence 2002 Aug 7 (pp. 403-410). Morgan Kaufmann Publishers Inc.

10. Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26.

11. Bengio, Yoshua, et al. A neural probabilistic language model // Journal of machine learning research 3.Feb (2003): 1137-1155.

12. Bruni, Elia, Nam-Khanh Tran, and Marco Baroni // Multimodal Distributional Semantics. J. Artif. Intell. Res.(JAIR) 49.1-47 (2014).

13. Levy, Omer, and Yoav Goldberg. Dependency-Based Word Embeddings. // ACL (2). 2014.

14. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. Glove: Global Vectors for Word Representation. // EMNLP. Vol. 14. 2014.

15. Moisl, Hermann. Cluster analysis for corpus linguistics. Vol. 66. Walter de Gruyter GmbH & Co KG, 2015.

16. McKeown, Kathleen R., et al. "Tracking and summarizing news on a daily basis with Columbia's Newsblaster." Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., 2002.

13. Материально-техническое обеспечение дисциплины

Для лекционных и практических занятий по темам дисциплины используется проектор и компьютеры с инструментальной средой программирования и выходом в сеть Интернет.

Авторы программы: _____/ Черняк Е.Л. /