Программа производственной практики

по направлению подготовки
45.03.03 «Фундаментальная и прикладная лингвистика»

Утверждена
Академическим советом ООП

Протокол № 5 от «26» июня 2017г.

Автор	А.Ю. Малафеев
Объем практики в з.е., кредитах	6
Продолжительность практики в академических часах, в т.ч. объем контактной работы в час.	228
Продолжительность практики в неделях	4
Курс	3
Вид практики	Производственная

1. **Цель и задачи производственной практики**

Целью проведения производственной практики является расширение и углубление знаний о возможных сферах приложения своих профессиональных знаний и приобретения практических навыков и компетенций в сфере фундаментальной и прикладной лингвистики, в условиях профессиональной деятельности в реальной организации.

Задачами практики могут являться следующие:

- приобретение навыков компьютеризированной разметки текстов;
- применение и развитие навыков программирования, особенно в сфере обработки текстов;
- .

2. **Способ проведения практики**

Способ проведения производственной практики – стационарный (в соответствии с пунктом 6.4 Образовательного стандарта по направлению подготовки 45.03.03 Фундаментальная и прикладная лингвистика, утвержденного Учёным советом НИУ ВШЭ, протокол от 30.01.2015 №1).

3. **Формы проведения практики**

Дискретная: путем выделения в календарном учебном графике непрерывного периода учебного времени для проведения каждого вида (совокупности видов) практики, в соответствии с учебным планом.

4. **Перечень планируемых результатов обучения при прохождении практики, соотнесенных с планируемыми результатами освоения образовательной программы**

Код компетенции	Формулировка компетенции	Содержание компетенции, которое формируется в ходе практики (дескрипторы освоения)	Профессиональные задачи, для решения которых требуется данная компетенция
УК-10	Способен осуществлять производственную или прикладную деятельность в международной среде	Студент способен работать в международных командах по решению различных лингвистических задач, а также участвовать в прикладных исследованиях совместно с зарубежными учеными, аспирантами и студентами	- коммуникация с членами исследовательской группы - проведение презентаций прикладного исследования - участие в проведении лингвистических экспериментов
ПК-11	Способен участвовать в создании представительных текстовых массивов, корпусов текстов, корпусов звучащей речи, мультимодальных корпусов, лингвистических и социолингвистических баз данных и пользоваться этими ресурсами	Студент может осуществлять профессиональную разметку текстов, умеет получать из различных источников, обрабатывать и анализировать текстовые данные.	- выполнение заданий руководителя практики: разметка текстов, сбор и анализ текстовых данных.
ПК-12	Способен проектировать системы анализа и синтеза естественного языка, анализа и синтеза мультимодальных языковых систем, в том числе лингвистических компонентов интеллектуальных и информационных электронных систем	Студент может проектировать и разрабатывать информационные электронные системы или отдельные модули этих систем, связанные с обработкой устной речи или письменных текстов на естественном языке	- выполнение заданий по написанию кода на языке программирования Питон или иных языках программирования
ПК-13	способен провести квалифицированное тестирование эффективности лингвистически ориентированного программного продукта	Студент умеет тестировать программные продукты, связанные с обработкой текстов на естественном языке	- выполнение заданий по тестированию программ, программных модулей и библиотек

5. Место практики в структуре ООП

Данная практика базируется на изучении следующих дисциплин:

- Теория языка
- Программирование

- Программирование для лингвистов
- Научно-исследовательский семинар
- Английский язык

Навыки, приобретенные во время прохождения производственной практики, могут быть применены студентами при подготовке выпускной квалификационной работы. Кроме того, навыки могут быть использованы при освоении следующих дисциплин учебного плана:

- Введение в нейронные сети и машинный перевод
- Практикум по разработке лингвистических систем
- Информационный поиск и извлечение данных
- Язык профессиональных коммуникаций

6. Содержание практики

п/п	Виды практической работы студента	Содержание деятельности	Формируемые компетенции
1	Экспертная разметка текстов	Студент выполняет задания руководителя практики по экспертной разметке текстов	УК-10, ПК-11
2	Создание компьютерных программ, программных модулей, компонентов и библиотек.	Студент выполняет задания руководителя практики по разработке компьютерных программ или программных компонентов на языке программирования Питон или иных доступных студенту языках программирования.	УК-10, ПК-12
3	Тестирование существующих (коммерческих) программных продуктов	Студент выполняет задания руководителя практики по тестированию существующих программных продуктов, используемых в Организации.	ПК-13

Практика проходит под руководством преподавателя факультета гуманитарных наук и сотрудника принимающей организации. Студенты могут проходить практику в организациях следующих типов:

- в IT-компаниях, активно использующих лингвистические технологии;
- в языковых школах, экзаменационных центрах, переводческих фирмах;
- в HR и информационных отделах организаций;
- в банках, бизнесе, государственных структурах;
- иных организациях, в которых востребованы знания и навыки в области фундаментальной и прикладной лингвистики.

Также возможно прохождение учебной практики в подразделениях или на информационных ресурсах ВШЭ (например, сайты образовательных программ и департаментов; медийные проекты, аффилированные с ВШЭ). Допускается распределенное (дискретное) прохождение практики в течение учебного года. В этом случае расчет длительности практики проводится по астрономическим часам из расчета 1 ЗЕ (кредит) = 38 академических часов = 26 астрономических часов. Часы прохождения практики не могут совпадать с часами обязательных учебных занятий по РУП. Заявление

на распределенную (дискретную) практику можно подавать в учебный офис в любое время до 1 июня текущего учебного года.

7. Отчетность по практике.

После прохождения практики (в течении 30 дней после её окончания) студенты представляют отчет о прохождении практики, в том числе (см. Приложения 1-3):

1. Дневник практики;
2. Отзыв руководителя практики.

8. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по практике

Примеры возможных формулировок заданий по практике:

Реализуйте наивный байесовский классификатор для автоматического определения общей тональности (положительная или отрицательная) англоязычных отзывов о фильмах. Для этого необходимо использовать следующие данные:

http://www.cs.cornell.edu/people/pabo/movie-review-data/

Это реальные отзывы пользователей imdb.com, прошедшие элементарную предобработку (токенизация, разбивка на предложения).

Задание 1. Составьте свой список англоязычной положительной и отрицательной лексики. Количество положительных и отрицательных слов может быть разным. Пример: good, bad, liked, hated. Объем списка - не менее 50 слов.

Задание 2. Для реализации классификатора необходимо написать следующие функции:

def extract_features(text, features):

Извлечение признаков. "Превращает" текст в вектор (список) признаков, т.е. чисел. Каждая позиция в списке - количество вхождений соответствующего слова из списка features в тексте text.

Вход: text (строка), features (список слов)

Выход: список целых чисел

def train_nbc(features, corpora):

Эта функция собственно "обучает" классификатор на корпусах. Необходимо посчитать параметры для двух моделей: модель положительных отзывов и модель отрицательных отзывов, на двух соответствующих корпусах. Другими словами, в одном списке должно быть два вложенных списка параметров, каждый длиной n, где n - количество признаков (слов). Каждый параметр считается так: количество вхождений соответствующего признака (слова) в соответствующем корпусе (положительном, когда строим первую модель, и отрицательном, когда строим вторую), делённое на суммарное количество вхождений всех признаков в данном корпусе.

Важно! При обучении желательно использовать сглаживание. Без сглаживания классификатор будет работать плохо. Стоит ему один раз увидеть в тексте признак, который ни разу не встречался в обучающей выборке для одного из классов, как классификатор "обнулит" весь результат для данного класса, независимо от того, какие остальные признаки были обнаружены в тексте. Чтобы

избежать этого досадного недоразумения, реализуем простое, но эффективное сглаживание "Плюс-1" (Add-1 smoothing). То есть для каждого корпуса мы сначала считаем реальные частоты для каждого признака, потом делаем сглаживание (все частоты прибавляются на 1), а уже после этого вычисляем параметры по формуле, приведённой в предыдущем абзаце.

Вход: features (список слов-признаков), corpora (список из двух корпусов, т.е. строк), classes (список из двух строк, обозначающих классы, напр. ['positive', 'negative'], в том же порядке, в котором идут "положительный" и "отрицательный" корпуса.

Выход: список списков (два списка чисел-параметров внутри одного списка)

def classify(text, features, classes, priors, Params):

Самая главная функция. Для каждого класса считает произведение вероятностей: вероятность класса (или prior) * вероятность каждого слова-признака для данного класса, возведённая в степень k, где k - абсолютная частота данного слова В АНАЛИЗИРУЕМОМ ТЕКСТЕ (а не в корпусе).

Вероятность класса в данной задаче одинакова для обоих классов и составляет 0.5 (т.к. у нас одинаковое количество текстов обоих классов).

Вероятность каждого слова-признака для каждого класса берётся из списка списков Params, который мы формировали ранее с помощью функции train_nbc.

После всех подсчётов функция "присваивает" анализируемому тексту тот класс, для которого произведение вероятностей получилось больше.

Вход: text (строка - анализируемый текст, который классификатор раньше не видел), features (список слов-признаков), classes (список из двух строк, обозначающих классы), priors (список из вероятностей каждого класса, в данном случае [0.5, 0.5]), Params (два списка чисел-параметров внутри одного списка).

Выход: строка, соответствующая названию класса, "выбранного" классификатором для анализируемого текста.

Вспомогательный код, который может идти после определения всех трёх функций:

```
corpus_pos = open('pos_train.txt', encoding='utf-8').read()
corpus_neg = open('neg_train.txt', encoding='utf-8').read()
corpora = corpus_pos, corpus_neg
classes = 'positive', 'negative'
Params = train_nbc(features, corpora, classes)
priors = 0.5, 0.5
text1 = 'The movie was horrible, it was absolutely awful.'
label1 = classify(text1, features, classes, priors, Params)
text2 = 'I really enjoyed the film, it was fantastic.'
label2 = classify(text2, features, classes, priors, Params)
print(label1, label2)  # ожидаемый вывод программы: negative positive
```

Прохождение производственной практики студентом оценивается по десятибалльной системе и учитывается при подведении итогов общей успеваемости студента.

Студенты, не выполнившие программы практик по уважительной причине, направляются на практику вторично, в свободное от учебы время.

Студенты, не выполнившие программы практик без уважительной причины или получившие по ее итогам неудовлетворительную оценку, считаются имеющими академическую задолженность. Она может быть ликвидирована в соответствии с Положением об организации промежуточной аттестации и текущего контроля успеваемости студентов Национального исследовательского университета "Высшая школа экономики".

Студенты, не выполнившие программы установленных видов практики на выпускных курсах, не допускаются к государственной итоговой аттестации.

9. Перечень учебной литературы и ресурсов сети «Интернет», необходимых для проведения практики

№ п/п	Наименование
Основная литература	
1	Математические основы машинного обучения и прогнозирования: Учебное пособие / Вьюгин В.В. - М.:МЦНМО, 2014 - 304 с.: ISBN 978-5-4439-2014-6 - Режим доступа: http://znanium.com/catalog/product/958689
Дополнительная литература	
2	Анализ данных и процессов: Учебное пособие / Барсегян А.А., Куприянов М.С., Холод И.И. - СПб:БХВ-Петербург, 2009 - 512 с. ISBN 978-5-9775-0368-6 - Режим доступа: http://znanium.com/catalog/product/350638

10. Перечень информационных технологий, используемых при проведении практики

В процессе прохождения учебной практики обучающиеся могут использовать информационные технологии, в том числе компьютерные программы, средства автоматизации проектирования и разработки программного обеспечения, Интернет - технологии и др.

Приложение 1. Форма титульного листа отчета о практике

Федеральное государственное автономное образовательное учреждение высшего образования
Национальный исследовательский университет
«Высшая школа экономики» -
Нижний Новгород

Факультет гуманитарных наук

О Т Ч Е Т

о прохождении _____ практики
(название практики)

в _____
(наименование организации)

Срок прохождения практики с _____ до _____

Студента _____
(фамилия, имя, отчество)
_____ курса _____ группы

Приложение 2. Форма отзыва руководителя

Отзыв руководителя практики от предприятия (организации) о работе студента

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

« _____» _____ 201___ г.

(подпись)

Приложение 3. Форма дневника практики

ДНЕВНИК ПРОИЗВОДСТВЕННОЙ ПРАКТИКИ СТУДЕНТА БАКАЛАВРИАТА НИУ ВШЭ – НИЖНИЙ НОВГОРОД

Студент _____

Курс, группа _____

Направляется на (в)

_____

(учреждение, предприятие)

_____

1. Календарные сроки практики

По учебному плану начало _____ конец _____

Дата прибытия на практику _____

Дата выбытия с места практики _____

2. Руководитель практики от университета.

Департамент _____ Уч. звание _____

Ф.И.О. _____

3. Руководитель практики от предприятия, учреждения.

Должность _____
Ф.И.О. _____

Ежедневные записи студента по практике

Дата	Описание работы, выполненной студентом	Отметки руководителя