“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. Ігоря Сікорського”

Кафедра системного програмування та спеціалізованих комп’ютерних систем

***ЗВІТ***

З переддипломної практики

студента IV курсу, групи КВ-41

**Курача Віктора Миколайовича**

Дата “ 16 ” березня 2018 р.

Підпис \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Назва підприємства: **НЦ “Кібербіонік Систематикс”**

Керівник практики від підприємства: Пономаренко О.

Дата “ 16 ” березня 2018 р.

Підпис \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |
| --- | --- |
| Керівник практики від  НТУУ “КПІ”  Наливайчук М. В. | Оцінка  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_)балів  Дата “ 16 ” березня 2018 р.  Підпис \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

Київ – 2018

**Зміст**

1. Призначення і завдання переддипломної практики
2. Характеристика підприємства, де відбувалась практика
3. Тема дипломного проекту. Постановка задачі. Аналіз завдання
4. Методи рішення та їх обґрунтування
5. Використана література

Додатки:

1. робоча програма практики;
2. анотація дипломного проекту українською, російською та англійською мовами;
3. технічне завдання дипломного проекту;
4. зміст пояснювальної записки дипломного проекту;
5. перший розділ дипломного проекту: «Аналіз існуючих рішень та обґрунтування теми дипломного проекту»

**Призначення і завдання переддипломної практики**

Переддипломна практика студентів є частиною процесу підготовки спеціалістів у вищих навчальних закладах і проводиться на оснащених відповідним чином базах навчальних закладів, а також на підприємствах і організаціях різних галузей господарства в області розробки і використання технічного та програмного забезпечення сучасної комп'ютерної техніки.

Призначенням практики є поглиблення та закріплення знань, отриманих у процесі навчання, набуття необхідних практичних та організаційних навичок роботи у сфері інформаційних технологій, підбір практичного матеріалу для написання дипломного проекту.

Переддипломна практика проводиться перед виконанням кваліфікаційної роботи. Робота студента в період переддипломної практики узгоджується з темою дипломного проекту. Під час проходження практики засвоюються нові технічні засоби та програмні прийоми, набуваються навички використання інструментарію, за допомогою якого виконується завдання дипломного проекту, будуються проміжні варіанти роботи, проводиться їх тестування та аналіз отриманих результатів.

Основні завдання, вирішення яких припадає на період проходження переддипломної практики:

1. Вибір теми дипломного проекту, постановка задачі.
2. Аналіз існуючих рішень та обґрунтування вибраної теми.
3. Вивчення структури та особливостей інструментарію.
4. Розробка технічного завдання дипломного проекту.
5. Вибір основних напрямків розв'язання завдання дипломного проекту.

**Характеристика підприємства, де відбувалась практика**

Навчальний центр «Кібербіонік Систематикс» («CyberBionic Systematics») проводить курси з навчання сучасним промисловим мовам програмування та супутнім технологіям, серед яких мови програмування Java, C#, Python, JavaScript, Ruby, платформа .NET, створення веб-порталів, управління базами даних тощо.

Крім того, навчальний центр проводить послуги з підвищення кваліфікації співробітників компаній, що працюють у галузі інформаційних технологій, їх сертифікації за стандартами Microsoft та послуги з IT-консалтингу.

Навчальний центр співпрацює з вузами (організовує семінари, тренінги, факультативи, надає спонсорську підтримку олімпіад та конкурсів, проводить програми підвищення кваліфікації для викладачів) та консультує своїх студентів з питань працевлаштування.

Компанія має два офіси: в Москві та в Києві в районі станції метро «Лівобережна».

**Тема дипломного проекту. Постановка задачі. Аналіз завдання**

Тема дипломного проекту: «Програмні засоби детектування слів іноземного походження в україномовних текстах».

Розробка проводиться мовою програмування Python 3.6 з використанням технологій машинного навчання.

У зв’язку з цим, робота над дипломним проектом передбачає вирішення наступних завдань:

1. Пошук розв’язків подібних задач для української та інших мов, їх аналіз.
2. Вивчення алгоритмів машинного навчання, вибір оптимального алгоритму за якістю результатів навчання та простотою реалізації.
3. Пошук розмічених текстів для проведення навчання системи.
4. Визначення характерних ознак слів іноземного походження.
5. Програмна реалізація алгоритму машинного навчання.
6. Тестування розробленого програмного продукту.
7. Оформлення документації дипломного проекту.

**Методи рішення та їх обґрунтування**

Машинне навчання – широкий підрозділ штучного інтелекту, що вивчає методи побудови алгоритмів, здатних до навчання. Розрізняють два типи навчання: *навчання по прецедентах*, або *індуктивне навчання*, що засноване на виявленні закономірностей в емпіричних даних, та *дедуктивне навчання*, що передбачає формалізацію знань експертів та їх перенос у комп’ютер у вигляді бази знань. Їх прийнято відносити до галузі експертних систем, тому терміни «машинне навчання» та «навчання по прецедентах» вважають синонімами.

Нехай дано скінченну множину прецедентів (об’єктів, ситуацій), по кожному з яких зібрано деякі дані. Дані про прецедент називають його описом. Сукупність усіх описів називається *навчальною вибіркою*. Задача навчання по прецедентах полягає в тому, що за цими даними необхідно виявити загальні закономірності, властиві не лише даній вибірці, а взагалі всім прецедентам, включаючи ті, які ще не спостерігались. Кажуть також про *відновлення залежності по емпіричних даних*.

Найбільш розповсюдженим способом опису прецедентів є *опис за ознаками*, який полягає в тому, що фіксується сукупність із *n* показників, що вимірюються в усіх прецедентів.

*Навчання з учителем* – найбільш поширений варіант навчання по прецедентах. Кожний прецедент є парою «об’єкт – відповідь». Задача полягає в знаходженні функціональної залежності відповідей від описань об’єктів та побудові алгоритму, що приймає на вході опис об’єкта і видає на виході відповідь.

Задача *класифікації* полягає в тому, що кількість допустимих об’єктів скінченна. Їх називають *мітками класів*. Клас – це множина всіх об’єктів із даним значеням мітки.

Задача *регресії* відрізняється тим, що допустимою відповіддю є дійсне число або числовий вектор.

Задача *ранжування* відрізняється тим, що відповіді треба отримати зразу на великій кількості об’єктів, а потім відсортувати їх по значеннях відповідей. Ця задача може зводитися до задач класифікації або регресії і часто використовується в інформаційному пошуку та аналізі текстів.

Задача *прогнозування* відрізняється тим, що об’єктами є відрізки часових рядів, які уриваються в момент, коли треба зробити прогноз на майбутнє. Для вирішення задач прогнозування часто можна використовувати методи регресії або класифікації.

[Обучение без учителя](http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B1%D0%B5%D0%B7_%D1%83%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D1%8F) (unsupervised learning). В этом случае ответы не задаются, и требуется искать зависимости между объектами.

* + Задача [кластеризации](http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F) (clustering) заключается в том, чтобы сгруппировать объекты в [кластеры](http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80&action=edit), используя данные о попарном сходстве объектов. Функционалы качества могут определяться по-разному, например, как отношение средних межкластерных и внутрикластерных расстояний.
  + Задача [поиска ассоциативных правил](http://www.machinelearning.ru/wiki/index.php?title=%D0%9F%D0%BE%D0%B8%D1%81%D0%BA_%D0%B0%D1%81%D1%81%D0%BE%D1%86%D0%B8%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D1%8B%D1%85_%D0%BF%D1%80%D0%B0%D0%B2%D0%B8%D0%BB&action=edit) (association rules learning). Исходные данные представляются в виде признаковых описаний. Требуется найти такие наборы признаков, и такие значения этих признаков, которые особенно часто (неслучайно часто) встречаются в признаковых описаниях объектов.
  + Задача [фильтрации выбросов](http://www.machinelearning.ru/wiki/index.php?title=%D0%A4%D0%B8%D0%BB%D1%8C%D1%82%D1%80%D0%B0%D1%86%D0%B8%D1%8F_%D0%B2%D1%8B%D0%B1%D1%80%D0%BE%D1%81%D0%BE%D0%B2&action=edit) (outliers detection) — обнаружение в обучающей выборке небольшого числа [нетипичных](http://www.machinelearning.ru/wiki/index.php?title=%D0%9D%D0%B5%D1%82%D0%B8%D0%BF%D0%B8%D1%87%D0%BD%D0%BE%D1%81%D1%82%D1%8C&action=edit) объектов. В некоторых приложениях их поиск является самоцелью (например, обнаружение мошенничества). В других приложениях эти объекты являются следствием ошибок в данных или неточности модели, то есть [шумом](http://www.machinelearning.ru/wiki/index.php?title=%D0%A8%D1%83%D0%BC&action=edit), мешающим настраивать модель, и должны быть удалены из выборки, см. также [робастные методы](http://www.machinelearning.ru/wiki/index.php?title=%D0%A0%D0%BE%D0%B1%D0%B0%D1%81%D1%82%D0%BD%D1%8B%D0%B5_%D0%BC%D0%B5%D1%82%D0%BE%D0%B4%D1%8B&action=edit) и [одноклассовая классификация](http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D0%B4%D0%BD%D0%BE%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%BE%D0%B2%D0%B0%D1%8F_%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D1%8F&action=edit).
  + Задача построения [доверительной области](http://www.machinelearning.ru/wiki/index.php?title=%D0%94%D0%BE%D0%B2%D0%B5%D1%80%D0%B8%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D0%BE%D0%B1%D0%BB%D0%B0%D1%81%D1%82%D1%8C&action=edit) (quantile estimation) — области минимального объёма с достаточно гладкой границей, содержащей заданную долю выборки.
  + Задача [сокращения размерности](http://www.machinelearning.ru/wiki/index.php?title=%D0%A1%D0%BE%D0%BA%D1%80%D0%B0%D1%89%D0%B5%D0%BD%D0%B8%D0%B5_%D1%80%D0%B0%D0%B7%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D1%81%D1%82%D0%B8&action=edit) (dimensionality reduction) заключается в том, чтобы по исходным признакам с помощью некоторых функций преобразования перейти к наименьшему числу новых признаков, не потеряв при этом никакой существенной информации об объектах выборки. В классе линейных преобразований наиболее известным примером является [метод главных компонент](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%B3%D0%BB%D0%B0%D0%B2%D0%BD%D1%8B%D1%85_%D0%BA%D0%BE%D0%BC%D0%BF%D0%BE%D0%BD%D0%B5%D0%BD%D1%82).
  + Задача [заполнения пропущенных значений](http://www.machinelearning.ru/wiki/index.php?title=%D0%97%D0%B0%D0%BF%D0%BE%D0%BB%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BF%D1%80%D0%BE%D0%BF%D1%83%D1%89%D0%B5%D0%BD%D0%BD%D1%8B%D1%85_%D0%B7%D0%BD%D0%B0%D1%87%D0%B5%D0%BD%D0%B8%D0%B9&action=edit) (missing values) — замена недостающих значений в матрице объекты–признаки их прогнозными значениями.
* [Частичное обучение](http://www.machinelearning.ru/wiki/index.php?title=%D0%A7%D0%B0%D1%81%D1%82%D0%B8%D1%87%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5) (semi-supervised learning) занимает промежуточное положение между обучением с учителем и без учителя. Каждый прецедент представляет собой пару «объект, ответ», но ответы известны только на части прецедентов. Пример прикладной задачи — автоматическая [рубрикация](http://www.machinelearning.ru/wiki/index.php?title=%D0%A0%D1%83%D0%B1%D1%80%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D1%8F_%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2&action=edit) большого количества текстов при условии, что некоторые из них уже отнесены к каким-то рубрикам.
* [Трансдуктивное обучение](http://www.machinelearning.ru/wiki/index.php?title=%D0%A2%D1%80%D0%B0%D0%BD%D1%81%D0%B4%D1%83%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5) (transductive learning). Дана конечная обучающая выборка прецедентов. Требуется по этим *частным* данным сделать предсказания отностительно других *частных* данных — [тестовой выборки](http://www.machinelearning.ru/wiki/index.php?title=%D0%A2%D0%B5%D1%81%D1%82%D0%BE%D0%B2%D0%B0%D1%8F_%D0%B2%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0). В отличие от стандартной постановки, здесь не требуется выявлять *общую* закономерность, поскольку известно, что новых тестовых прецедентов не будет. С другой стороны, появляется возможность улучшить качество предсказаний за счёт анализа всей тестовой выборки целиком, например, путём её кластеризации. Во многих приложениях трансдуктивное обучение практически не отличается от частичного обучения.
* [Обучение с подкреплением](http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81_%D0%BF%D0%BE%D0%B4%D0%BA%D1%80%D0%B5%D0%BF%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5%D0%BC) (reinforcement learning). Роль объектов играют пары «ситуация, принятое решение», ответами являются значения функционала качества, характеризующего правильность принятых решений (реакцию среды). Как и в задачах прогнозирования, здесь существенную роль играет фактор времени. Примеры прикладных задач: формирование инвестиционных стратегий, автоматическое управление технологическими процессами, самообучение роботов, и т.д.
* [Динамическое обучение](http://www.machinelearning.ru/wiki/index.php?title=%D0%94%D0%B8%D0%BD%D0%B0%D0%BC%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5&action=edit) (online learning) может быть как обучением с учителем, так и без учителя. Специфика в том, что прецеденты поступают потоком. Требуется немедленно принимать решение по каждому прецеденту и одновременно доучивать модель зависимости с учётом новых прецедентов. Как и в задачах прогнозирования, здесь существенную роль играет фактор времени.
* [Активное обучение](http://www.machinelearning.ru/wiki/index.php?title=%D0%90%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5&action=edit) (active learning) отличается тем, что обучаемый имеет возможность самостоятельно назначать следующий прецедент, который станет известен. См. также [Планирование экспериментов](http://www.machinelearning.ru/wiki/index.php?title=%D0%9F%D0%BB%D0%B0%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D1%8D%D0%BA%D1%81%D0%BF%D0%B5%D1%80%D0%B8%D0%BC%D0%B5%D0%BD%D1%82%D0%BE%D0%B2&action=edit).
* [Метаобучение](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D0%B0%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5&action=edit) (meta-learning или learning-to-learn) отличается тем, что прецедентами являются ранее решённые задачи обучения. Требуется определить, какие из используемых в них [эвристик](http://www.machinelearning.ru/wiki/index.php?title=%D0%AD%D0%B2%D1%80%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0&action=edit) работают более эффективно. Конечная цель — обеспечить постоянное автоматическое совершенствование [алгоритма](http://www.machinelearning.ru/wiki/index.php?title=%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC) обучения с течением времени.
  + [Многозадачное обучение](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5&action=edit) (multi-task learning). Набор взаимосвязанных или схожих задач обучения решается одновременно, с помощью различных алгоритмов обучения, имеющих схожее внутренне представление. Информация о сходстве задач между собой позволяет более эффективно совершенствовать алгоритм обучения и повышать качество решения основной задачи.
  + [Индуктивный перенос](http://www.machinelearning.ru/wiki/index.php?title=%D0%98%D0%BD%D0%B4%D1%83%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D1%8B%D0%B9_%D0%BF%D0%B5%D1%80%D0%B5%D0%BD%D0%BE%D1%81&action=edit) (inductive transfer). Опыт решения отдельных *частных* задач обучения по прецедентам переносится на решение последующих *частных* задач обучения. Для формализации и сохранения этого опыта применяются реляционные или иерархические структуры представления знаний.
  + Иногда к метаобучению ошибочно относят построение [алгоритмических композиций](http://www.machinelearning.ru/wiki/index.php?title=%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BA%D0%BE%D0%BC%D0%BF%D0%BE%D0%B7%D0%B8%D1%86%D0%B8%D1%8F&action=edit), в частности, [бустинг](http://www.machinelearning.ru/wiki/index.php?title=%D0%91%D1%83%D1%81%D1%82%D0%B8%D0%BD%D0%B3); однако в композициях несколько алгоритмов решают *одну и ту же* задачу, тогда как метаобучение предполагает, что решается много разных задач.

**Використана література**

1. Бибик С. П. Словник іншомовних слів: тлумачення, словотворення та слововживання / Бибик С. П., Сюта Г. М. За ред. С. Я. Єрмоленко // Харків: Фоліо. – 2006. – 623 с.

2. Машинное обучение [Електронний ресурс]. – Режим доступу: http://machinelearning.ru/. – Назва з екрану. – Дата звернення: 13.03.2018.

3. Donald Michie, D.J. Spiegelhalter, C.C. Taylor. Machine learning, neural and statistical classification. Ellis Horwood Upper Saddle River, NJ, USA. 1994.

4. Nils J. Nilson. Introduction To Machine Learning. Robotics Laboratory, Department Of Computer Science, Stanford Univercity, Stanford, CA, USA. 2005.