**Задание на выпускную квалификационную работу бакалавра ИВТ по профилю АСУ и план подготовки ВКР к защите на 2020/2021 учебный год.**

|  |  |
| --- | --- |
| Научный руководитель | Студент |
| Будаева А.А. |  |

1. **Название и содержательная постановка задачи.**

*Тема работы:* **«Разработка системы кластеризации документов на базе методов машинного обучения»**

*Содержательная постановка задачи:*

Имеется фиксированное множество документов. Требуется разбить заданное множество документов на группы в соответствии с их семантической схожестью.

Предусмотреть 2 режима работы:

1. автоматическое выявление групп документов
2. выбор одной группы документа (из заранее определенных групп) на основании его содержания.

В качестве критериев схожести документов при кластеризации используется мера TF-IDF

*Область применения*:

* классификация статей на конференции
* классификация документов по категориям
* и др.

1. **Обозначения и формальная постановка задачи**

Обозначения:

*D* – исходное множество документов коллекции,

*di* ∈ *D* – *i*-й документ множества *D*,

*tj* ∈ *di* – *j*-е слово документа *di*,

 – число вхождений *tj*-го слова в документ *di*,

 – общее число слов в документе *di*, 

 – частота *tj*-го слова (term frequency) в документе *di*,

 – обратная частота документа (inverse document frequency) – инверсия частоты, с которой *tj*-ое слово встречается в документах коллекции

– число документов *D*, в которых встречается слово *tj* (когда )

 – булева переменная, равная 1 – если *di*-й документ относится к таксону *pk* и 0 – в противном случае,

 – оценка важности слова в контексте документа (мера TF-IDF)

- евклидово расстояние между *i*-м и *l*-м документами на основании меры TF-IDF



**Цель** – суммарное расстояния между документами, попавшими в один таксон, должно быть минимальным.

****

1. **План подготовки ВКР к защите**
   1. **Ноябрь 2020:** аналитический обзор методов решения поставленной задачи и существующих пакетов программ, способных её решать.
   2. **Декабрь 2020**: Выбор алгоритма и его программная реализация. В конце месяца – отчет на кафедре.
   3. **Январь 2021**: Набор статистики, построение экспериментальных графиков. Статистическая оценка качества найденных в разделе 3.3 аналитических зависимостей: коэффициенты корреляции, дисперсия, доверительные интервалы и т.п.
   4. **Февраль2021**: Добавление базы данных и дружественного интерфейса в созданный программный пакет. В конце месяца – отчет на кафедре.
   5. **Март - апрель 2021:** Оформление ВКР, подготовка презентации и доклада.
   6. **Май 2021**: Предзащита и внесение корректив в доклад и презентацию (если необходимо).
   7. **Июнь 2021:** защита ВКР.