**Задание на выполнение исследований по дисциплине: «Информационные процессы в социотехнических системах»**

Тема работы: Экспериментальное исследование точности работы алгоритмов кластеризации текстов, написанных на естественном языке.

**I. Исходные данные:**

Ответственный студент от группы: *Калинин В.Н.*

Перечень алгоритмов для подгруппы:

- Алгоритм 1: *Agglomerative Clustering*;

- Алгоритм 2: *Self-Organizing Maps*;

- Алгоритм 3: *CACTUS*.

Датасеты текстов для подгруппы:

- Датасет 1: «*SELHOZ*» (200 больших, средних и коротких текстов);

- Датасет 2: «*PYTTYR*» (200 больших, средних и коротких текстов);

- Датасет 3: «*CULTUR*» (200 больших, средних и коротких текстов).

**II. Требования**

Отчет необходимо оформить по требованиям к оформлению работ студентов-бакалавров на кафедре КБ-8.

Не допустимо использовать полностью готовые решения (код, готовая библиотека), алгоритм программы должен быть написан самим студентом (за исключением сложных алгоритмов (в частности нейросетей), требующих фундаментальных знаний в данной области).

Допустимые языки программирования для решения задачи: *С++, C#, Python, Java*.

Отчет необходимо предоставить до: *Сентябрь 02, 2022*.

Структура отчета:

- Титульный лист

- Введение

- Основная часть

1. Описание исследуемых алгоритмов, блок-схема алгоритма, код;

2. Выполнение задачи, исследование точности алгоритмов на выданных датасетах. (графики, метрики, точность)

3. Дополнительная часть

- Выводы

- Список используемых источников

- Приложения

**III. Постановка задачи**

Векторная модель текста создается на основе:

1. Лексического подхода с использованием:

- "Обычной" частотной модели TF-IDF;

- Бинарной модели вхождения терминов в документ;

- Модели в которой TF-нормализованная частота, а IDF - "обычная" обратная частота;

- Модели в которой TF - нормализованная частота, IDF - сглаженная обратная частота.

При лексическом подходе использовать:

a) только частоты вхождения слов,

b) слова и n-граммы,

c) ассоциативно - семантические классы для терминов входящих в документы, полученные с использованием алгоритма ICAN.

2. Семантического подхода с использованием алгоритма ICAN.

**IV. Задача**

Исследовать для выданного перечня алгоритмов (**3 алгоритма**) точность их работы на множестве калиброванных по темам на основе экспертной оценки текстов (**9 датасетов** текстов), также ранее выданных.

**Важно!!!** Каждый **датасет – 200 текстов**! Не допускается уменьшать количество текстов, модифицировать их или объединять в общую тему!!!

При исследовании точности работы алгоритмов построить графики зависимостей точности от метрик, используемых в алгоритмах кластеризации (задаваемое число кластеров, мера сходства векторов при отнесении к одному классу и т.д.).

Точность работы алгоритмов оценивать на основе вхождения размеченных векторов в различные классы.

**V. Дополнительная часть**

Дополнительная часть **для тех, кто претендует на оценку отлично** по дисциплине: «Информационные процессы в социотехнических системах».

1. Для выданного корпуса текстов провести тематическое моделирование (статистический метод для поиска тем в коллекции документов) на основе Latent semantic analysis (LSA), а также Probabilistic Latent Semantic Analysis (PLSA) и Latent Dirichlet Allocation (LDA).

2 Для выданного корпуса текстов провести определение эмоциональной окраски.

3. Выбрать на новостном портале или социальной сети новость любой тематики с комментариями посетителей и провести оценку эмоциональной окраски.