

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СУМСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

ІНФОРМАТИКА, МАТЕМАТИКА,
АВТОМАТИКА

ІМА :: 2016

**МАТЕРІАЛИ
та програма**

НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

(Суми, 18–22 квітня 2016 року)



Суми
Сумський державний університет
2016

Применение косинусной меры для классификации статей по УДК

Колесниченко А.Л., студент

Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия

Одним из наиболее эффективных методов классификации текстов является метод косинусной меры. Он заключается в нахождении минимального угла между векторами, соответствующими темам, и вектором, соответствующим тексту. Векторы формируются из числа встреч каждого ключевого слова в тексте. Из-за древовидной структуры УДК, было решено принимать в качестве результата не одну, а две наиболее вероятных темы.



Рисунок 1 – Результаты применения метода косинусной меры.

После проведения классификации выполняется анализ ошибок (рис. 2), который показывает, какое процентное количество текстов для каждой темы было ошибочно отнесено к другим.

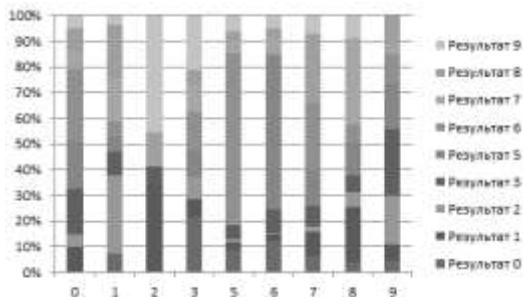


Рисунок 2 – Анализ ошибок.

Руководитель: Романов А.Ю., старший преподаватель