ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

SELF-STUDY

Đề tài: Spam Text Classification

Mentor: Online Course
Student: Lại Khánh Hoàng

Tp. Hồ Chí Minh

Mục lục

1 N-gram

Định nghĩa: các chuỗi có N từ liên tiếp
Example: "please turn your page"

- Unigram: please, turn, your, page
- Bigram: please turn, turn your, your page
- Trigram: please turn your, turn your page

2 Inverse Document Frequency (IDF)

Với mỗi từ $word$, $IDF(word) = \log \frac{N+1}{n_{word}+1} + 1$.
Trong đó, N là tổng số hàng dữ liệu và n_{word} là số hàng chứa từ $word$.

3 TF-IDF

$$\text{TF-IDF} = \text{TF} * \text{IDF}$$