CS246: Twitter Purification

Wen Shi, Zijun Xue, Jennifer Zhang

June 2014

Abstract

Introduction

Motivation

Overview

Implementation

Dictionary

TODO: Jennifer

Single word correction

Abbreviations: single word

TODO: Jennifer

Squeeze

TODO: Zijun

Edit distance

TODO: Zijun

Phonetic candidates: Soundex

TODO: Shi Wen

Phonetic candidates: Metaphone

TODO: Shi Wen

Letter similarity: Viterbi

TODO: Jennifer

Word frequency scaling

TODO: Jennifer

Abbreviations: phrase

Bigram correction

TODO: add some subsections here?

Discussion and Evaluation

Single word results

Bigram results

Future work

TODO: listing any more aspects where it didn't work well

Word tokenization

Obscure vocabulary

Proper nouns and acronyms

Abbreviations

Dictionary

Metadata: capitalization, emoticons, hashtags, usernames