Классификация записей из микроблогов с помощью Википедии

Абишев Т. М.
545 группа, математико-механический факультет, СПбГУ

Научный руководитель: Барашев Д. В.
Доцент, математико-механический факультет, СПбГУ

Конференция СПИСОК
СПбГУ, Санкт-Петербург, апрель 2012

Введение

- Данных всё больше
- Микроблоги как пример таких данных
- ...Которые необходимо классифицировать
  - зачем?
- И которые имеют особенности
  - какие?

Постановка задачи

- Построить классификатор сообщений из микроблогов
- Который будет использовать
  - википедию
  - контекст
- Результатом классификации хотелось бы видеть категории из википедии

Идеи

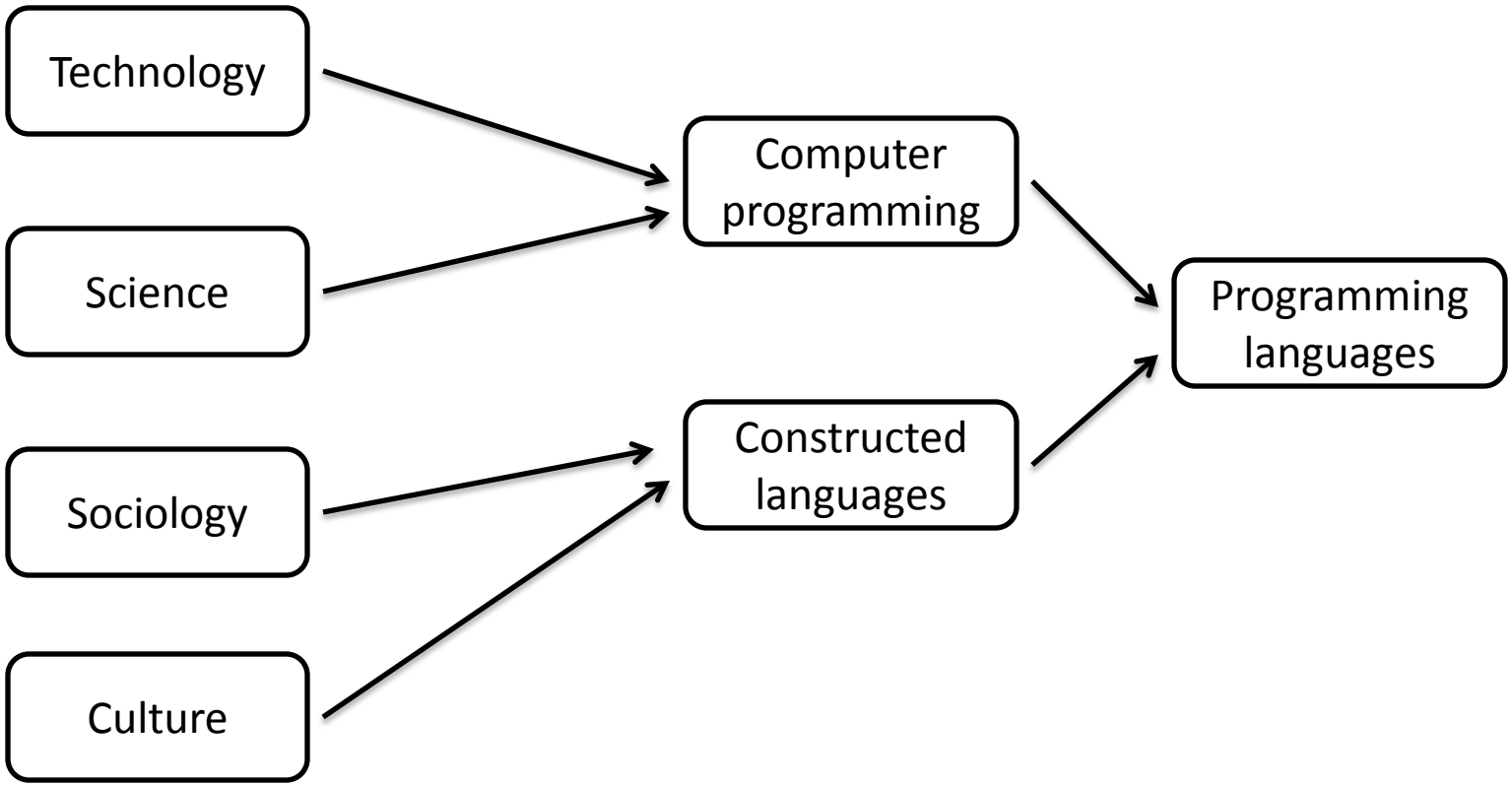
- Википедия, как источник помощи при классификации
  - категории
  - ссылочная структура
- Контекст
  - о чем еще пишет автор данного сообщения?

Проблемы (1)

- Размер одного сообщения
- Размер английской википедии
- Структура категорий
  - совсем не дерево
  - и даже не ацикличный граф

Проблемы (2)

- *Programming languages*?



Идеи для классификации одного сообщения

- Будем использовать подсчитанную статистику слов по которым делаются переходы на страницы википедии
- Для каждого слова сообщения найдем одну-две наиболее вероятных статьи на которые эти слова ведут
- Попробуем найти категорию википедии включающую наибольшую долю найденных выше статей

Идеи для классификации одного сообщения с учетом контекста

- Будем использовать найденные ранее категории, как features для кластеризации
- Кластеризуем другие сообщения автора
- Сопоставим каждому кластеру объединяющую категорию/категории
- Найдем к какому из кластеров относится сообщение для классификации

Текущие результаты (1)

- Написан подсчет/извлечение различных статистик/данных из википедии:
  - граф категорий
  - данные о перенаправлениях
  - статистика слов
  - ссылочная статистика
- На основе этих данных написан простейший алгоритм оценки того насколько много пользователь пишет о программировании

Текущие результаты (2)

- Наивный алгоритм
  - Подсчитана статистика по всем словам насколько часто они ведут на страницу принадлежащую (прямо или косвенно) категории программирование
  - rank(word) = count(word) / count(all words in anchors)
  - rank(message) = sum of rank(words)
  - rank(user) = sum of rank(message) / count of messages.

Текущие результаты (3)

- Программисты
  - sstephenson: 16.15
  - joehewitt: 2.05
  - joestump: 4.15
- Не программисты
  - katyperry: 0.30
  - ladygaga: 0.92
  - BarackObama: 0.20

Дальнейшее развитие

- Реализация предложенных идей
- Эксперименты