Проект “ANTIDICT”

База морфологических структур, допустимых в современном русском языке

Студенты:
Алексей Доркин
Владислава Смирнова
Антон Вахранев

Куратор: Варвара Магомедова

Что мы успели сделать

– Начали разделять слова по группам по определенному признаку (например, по общим словообразовательным признакам)

– Обработали и разметили словарь Дьякова

– Выкачали словарь Зализняка со словоформами

– Написали классификатор для заимствований, обучили модель

Группируем слова

накачаныйочкарик прекрасныеройдёный непосоветуют любойпредмет теперьпоменялось болельщикоспидигонщик	репортажечка геймерюга быдлячестве степанскую громыхалки жизнеоформляющих кивибоем редбуловых асогласна тиражир омедведюсь набухтелся	ловкаческой ребенковская перехрустит лунограддцы фидошников манхетенской	фсячности афтобус пионэром бидилы трэскат подлечица чумазег обзатце пережки маральнае жывотные
дезастером ресэмплинг тролбейн ргб трекеров		ляляка мимишном щупака свиняки милошных	

Группируем слова

антиудаизмом	электрагенератары	новогодих	оечка
неправославное	географисеских	создвездия	суперартмаркет
антифосфатные	переводишьс	подтведили	мармонам
неоценившему	справяю	бершадського	четров
некапиталистическим	разобразным	обяснило	катинками
гиперкомпенсировать	кажеться	мудчины	родоночальник
киноальянс	документацыи	имланты	эспресионистская
горетуристами	сковыжималку	чужародная	галаграма
видеоклипам	отсутстви	джнисы	эфективное
фотослайдами	компъютер	накуриваються	псхиушки
китчевость	фотоапарата	сверхъественого	
схемку	метериться	компенисрующие	
впариванию	неизменёных	рождествеская	
выдеорепортажем			
тромбозникам			

Размеченные данные из Дьякова

__label__neo юникс
__label__neo юнилевер
__label__neo юнилэтаризм
__label__neo юнион
__label__neo юнионизм
__label__neo юнионист
__label__neo юнионисты
__label__neo юниорский
__label__neo юнипол
__label__neo юнисеф
__label__neo юнискан
__label__neo юнит
__label__neo юнитард
__label__neo юнити

__label__dict лидируем
__label__dict воздуховоде
__label__dict гугенотке
__label__dict развёрстывавшуюся
__label__dict взъедавшееся
__label__dict перебинтовывающим
__label__dict прирастится
__label__dict нимфоманку
__label__dict покормившийся
__label__dict натуралисту
__label__dict учтивее
__label__dict поваловей

T3

https://docs.google.com/document/d/1fpn72q8bKqhFnCaTmbqGZEtWS6WcXIP9WC1VpviAGEc/edit#

Репозиторий

https://github.com/wksmirnowa/antidict

О чем проект (слайд на случай, если кто-то забыл)

– изучение необычных слов, которые можно встретить в интернете (50 млн словоформ из ГИКРЯ)

– классификация слов по большим категориям (например, заимствования, экспрессивные формы, слова со случайными и с намеренными ошибками и искажениями)

– классификация слов по подкатегориям (например, заимствования, слова с заимствованными морфемами, слова, образованные от русских корней)

– морфологический анализ слов