Aris Kosmopoulos «Large Scale Hierarchical Text Classification»

Остапец Андрей

28 октября 2015 г.

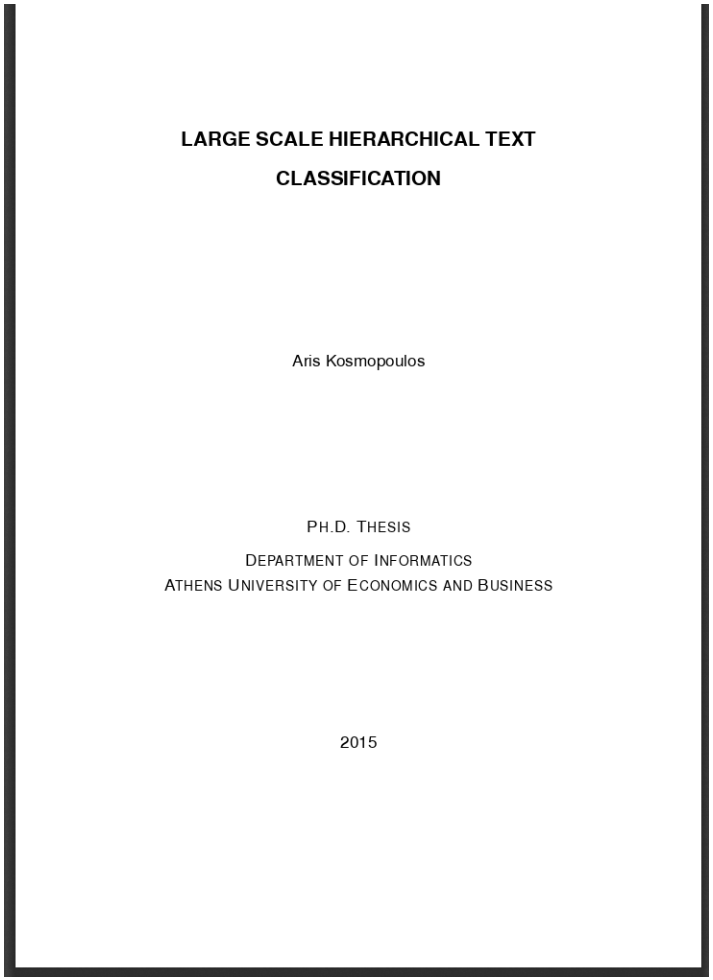
Aris Kosmopoulos

Биография:

- Athens University of Economics and Business, Master, Computers Science 2001 – 2007
- Athens University of Economics and Business, Doctor of Philosophy (PhD) 2008 – 2015
- BioASQ: October 2012

Aris Kosmopoulos

Ph.D. Thesis



LARGE SCALE HIERARCHICAL TEXT CLASSIFICATION

Aris Kosmopoulos

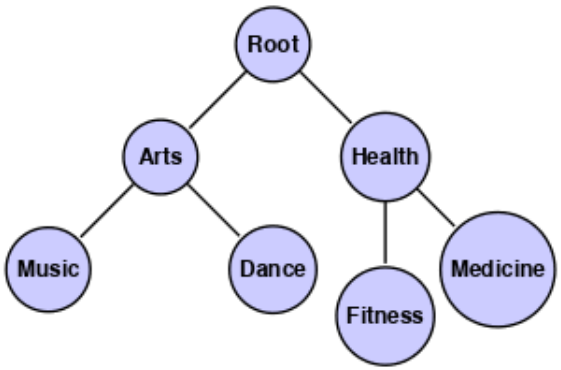
PH.D. THESIS

DEPARTMENT OF INFORMATICS
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

2015

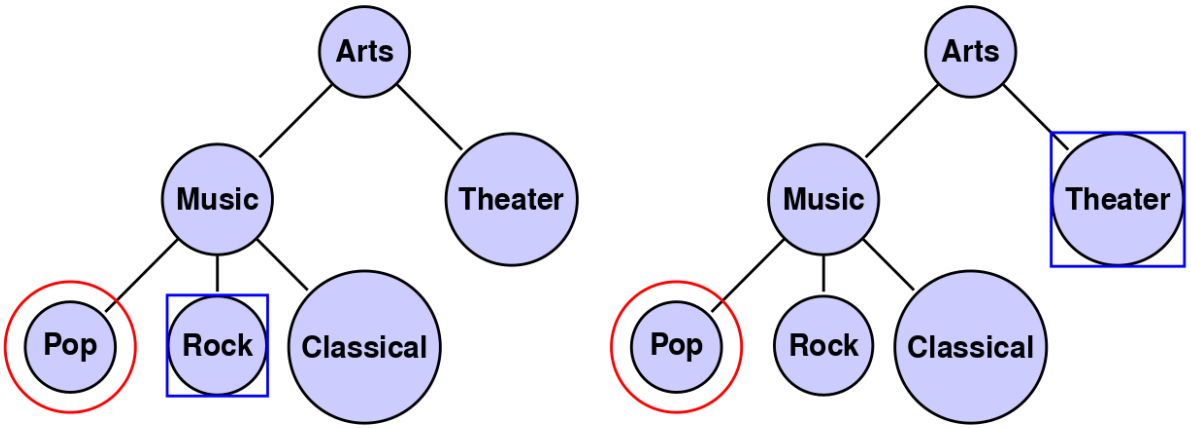
Таксономия

- Иерархия категорий представлена в форме отношений между вершинами
- Тысячи категорий
- Сотни тысяч или даже миллионы документов
- Каждый документ может принадлежать одной или нескольким категориям
- Некоторые категории могут иметь очень мало объектов в обучающей выборке



Проблема

«Плоские» меры качества игнорируют степень ошибки



Каждый объект принадлежит строго одной категории

- Accuracy $= \frac{\sum_{i=1}^{|C|} TP_i}{|D|}$
- Macro-precision $= \frac{\sum_{i=1}^{|C|} precision_i}{|C|}$, $precision_i = \frac{TP_i}{TP_i + FP_i}$
- Macro-recall $= \frac{\sum_{i=1}^{|C|} recall_i}{|C|}$, $recall_i = \frac{TP_i}{TP_i + FN_i}$
- Macro-$F_1 = \frac{2 \cdot \text{Macro-precision} \cdot \text{Macro-recall}}{\text{Macro-precision} + \text{Macro-recall}}$
- где C - это категории, D - это тестовые документы

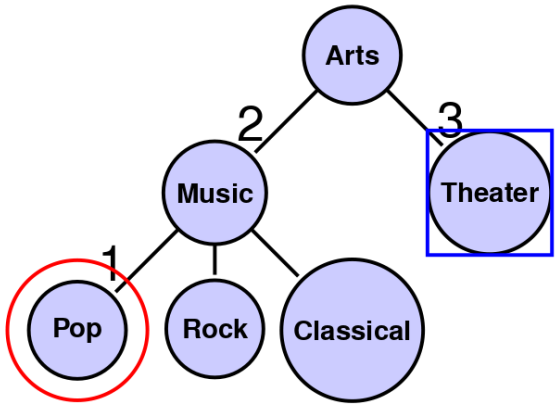
Каждый объект может принадлежать нескольким категориям

- Accuracy $= \frac{\sum_{i=1}^{|D|} \frac{|P_i \bigcap T_i|}{|P_i \bigcup T_i|}}{|D|}$
- Precision

  Macro-precision $= \frac{\sum_{i=1}^{|C|} precision_i}{|C|}$, Micro $= \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|}(TP_i+FP_i)}$
- Recall

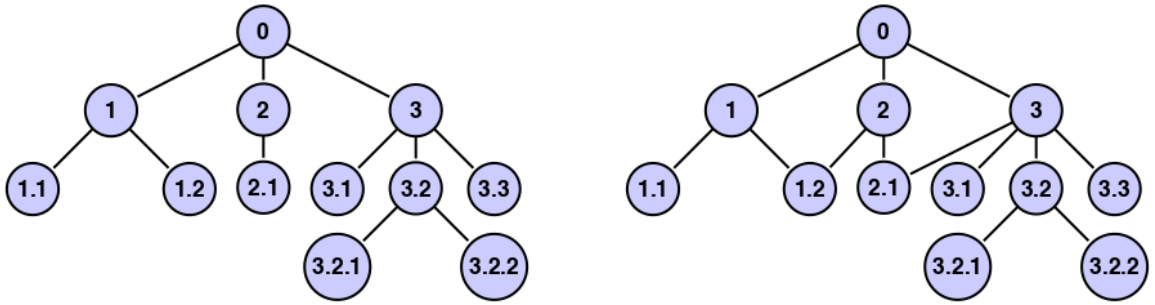
  Macro $= \frac{\sum_{i=1}^{|C|} recall_i}{|C|}$, Micro $= \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|}(TP_i+FN_i)}$
- Macro версии «уравнивают» все классы, micro версии отдают «больший» вес частотным классам
- где C - это категории, D - это тестовые документы, P_i - множество предсказанных категорий, T_i -множество истинных категорий

Простейшее использование иерархии



- Tree-induced error = $\frac{\sum_{i=1}^{D} \text{Shortest-path}(P_i, T_i)}{|D|}$
- D - тестовые документы
- P_i - категория, куда документ был классифицирован
- T_i - истинная категория документа
- Подходит только для случаев, когда иерархия категорий - это дерево и каждый документ принадлежит одной категории

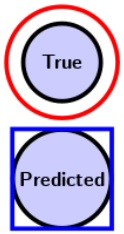
Что хотелось бы?

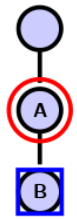


(a) Tree (b) DAG

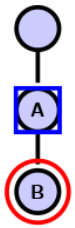
- Универсальное решение, когда документ может принадлежать строго одной категории, а может нескольким
- Различные виды таксономии (дерево, направленный ациклический граф, $\cdots$)
- Возможна классификация не только в листья, но и во внутренние узлы

Важные случаи

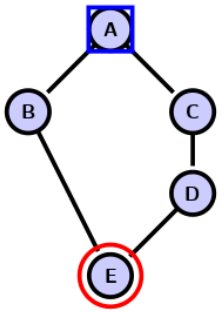




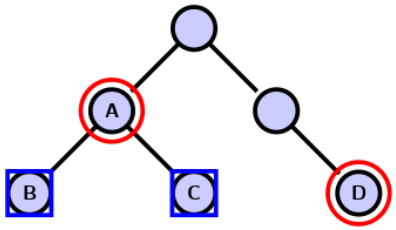
Over-specialization



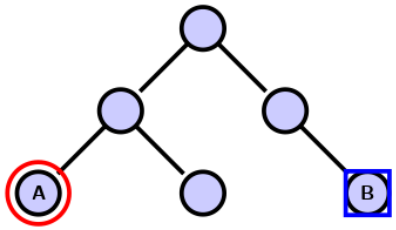
Under-specialization



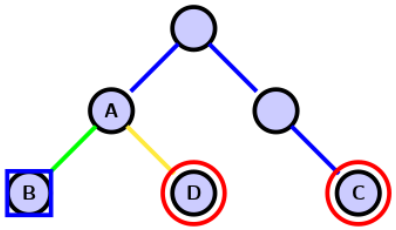
Alternative paths



Pairing problem



Long distance problem

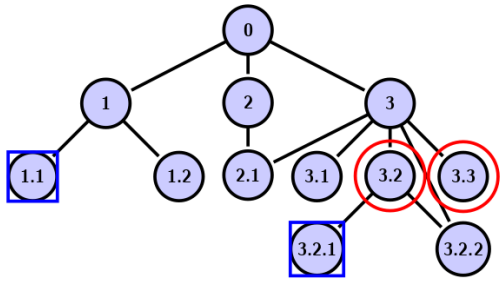


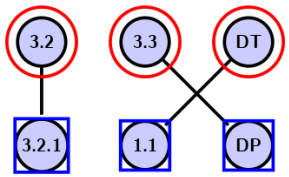
Multiple path count

2 возможных общих подхода

- Меры качества, основанные на попарном сравнении (pair-based measures)
  - Каждой предсказанной категории ставится истинная категория и наоборот
  - Вычисляется расстояние между каждой парой
  - Итоговая мера качества вычисляется на основе вычисленных расстояний
- Меры качества, основанные на сравнении множеств (set-based measures)
  - Вычисляются два множества (Y и $\hat{Y}$) из истинных и предсказанных категорий
  - Каждое множество расширяется (Y_{aug} и $\hat{Y}_{aug}$), используя дополнительные узлы из иерархии
  - Итоговая мера качества вычисляется на основе (Y_{aug} и $\hat{Y}_{aug}$)

Меры качества, основанные на попарном сравнении

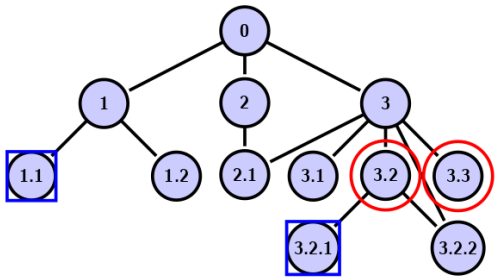


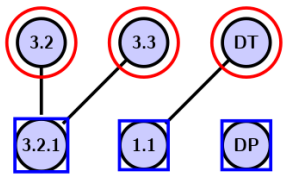


- В скольких парах может участвовать одна вершина?
- Использовать или нет вершины (Default Predict, Default True)?
- Какой порог выбрать для соединения с вершинами (Default Predict, Default True)?

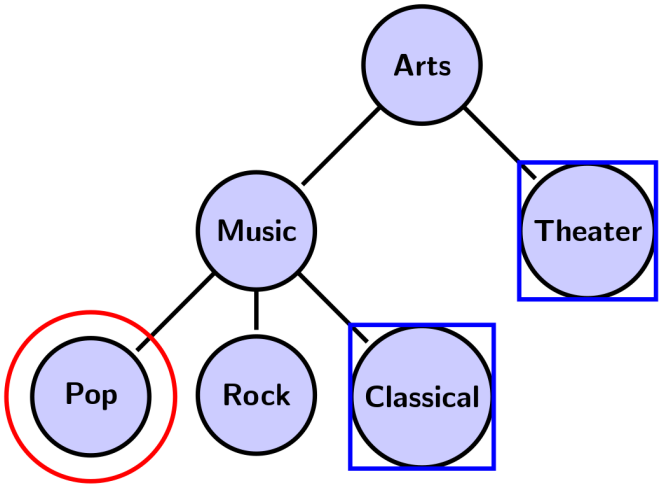
Multi-label Graph Induced Accuracy

- Каждой вершине (предсказанной и истинной) ставится в соответствие ближайшая вершина из другого класса
- Разрешено ставить в соответствие вершины по умолчанию
- Порог для соединения с ними - параметр.
- $1 - \frac{\text{Сумма расстояний между парами}}{|(P \bigcup T) \backslash (P \bigcap T)| \cdot D_{max}}$



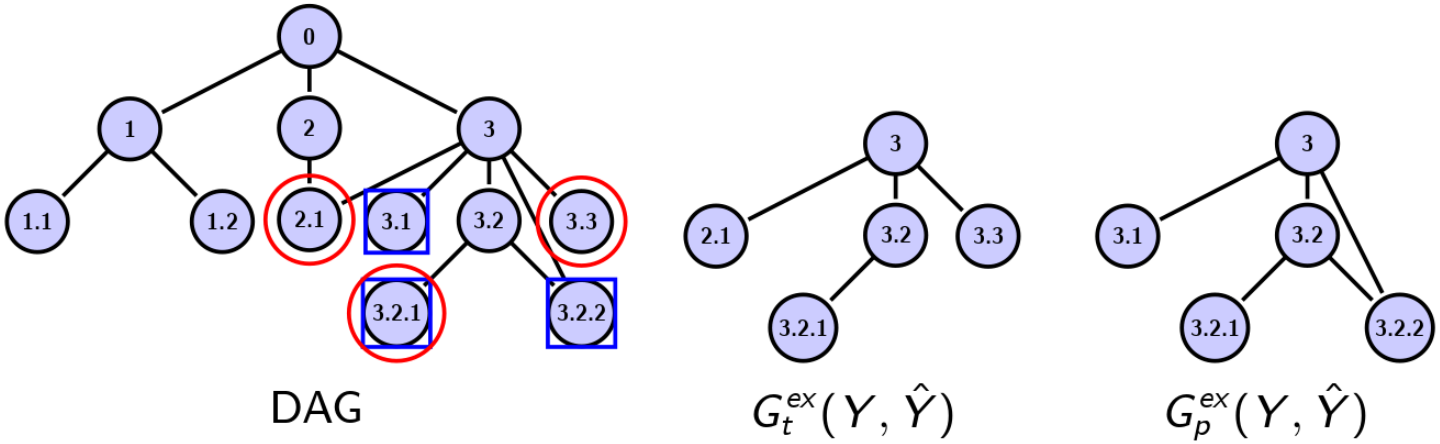


Меры качества, основанные на сравнении множеств



- $Y = \{\text{Pop}\}$, $\hat{Y} = \{\text{Classical, Theater}\}$
- $Y_{aug} = \{\text{Pop, Music, Arts}\}$, $\hat{Y}_{aug} = \{\text{Classical, Theater, Music, Arts}\}$
- $P_H = \frac{|\hat{Y}_{aug} \bigcap Y_{aug}|}{|\hat{Y}_{aug}|}, R_H = \frac{|\hat{Y}_{aug} \bigcap Y_{aug}|}{|Y_{aug}|}$

Least Common Ancestok F_1



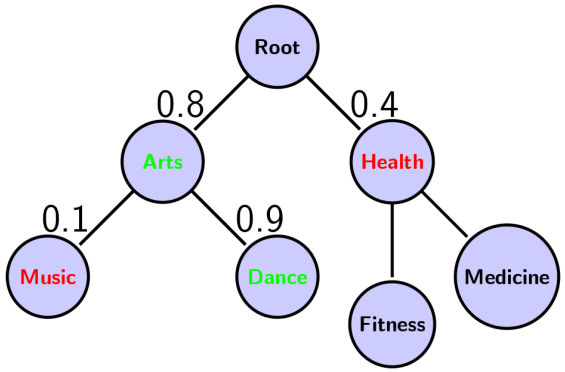
- Каждой вершине (предсказанной и истинной) ставится в соответствие ближайшая вершина из другого класса
- Создаются 2 графа: $G_t^{ex}(Y, \hat{Y})$, $G_p^{ex}(Y, \hat{Y})$
- Вычисляются P_H и R_H используя вершины из $G_t^{ex}(Y, \hat{Y})$ как Y_{aug}, вершины из $G_p^{ex}(Y, \hat{Y})$ как $\hat{Y}_{aug}$

Итоговое сравнение

	Single-label	Multi-label
Tree hierarchy	Any	LCA or MGIA
DAG hierarchy	LCA or MGIA	LCA or MGIA

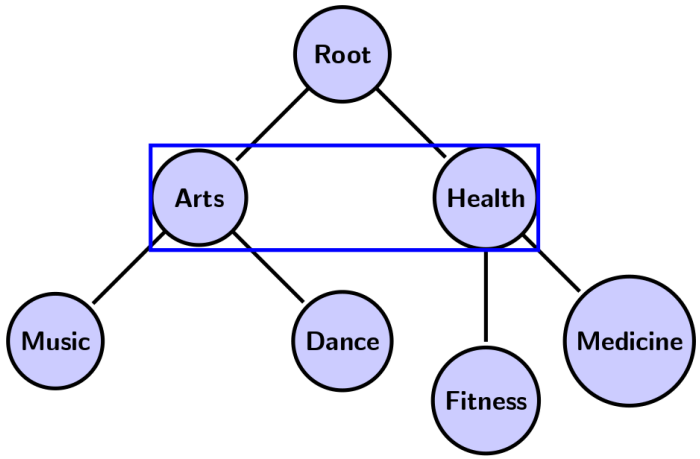
	pair-based		set-based measures	
	Tree Induced Error	MGIA	F_H	F_{LCA}
Alternative Paths	-	+	-	+
Over-specialization	+	+	+	+
Under-specialization	+	+	+	+
Pairing Problem	-	+	+	+
Long Distance Problem	+	+	+	+
Multiple Path Count	-	-	+	+

Каскад классификаторов



- Бинарный классификатор обучается для каждого узла, кроме корневого
- Классификатор в узле Arts использует:
  - Документы категорий *Music* и *Dance* как позитивные
  - Документы категорий *Fitness* и *Medicine* как негативные
- Каждый тестовый документ проходит по иерархии от корня до листа, выбирая каждый раз наиболее вероятное направление

Уменьшение размерности на верхних уровнях



- Bag of words: каждое слово - это признак
- Уменьшение размерности признаков:
  - Может обрабатывать больше документов
  - Уменьшение переобучения
- Размер словаря значительно увеличивается при подъеме по иерархии

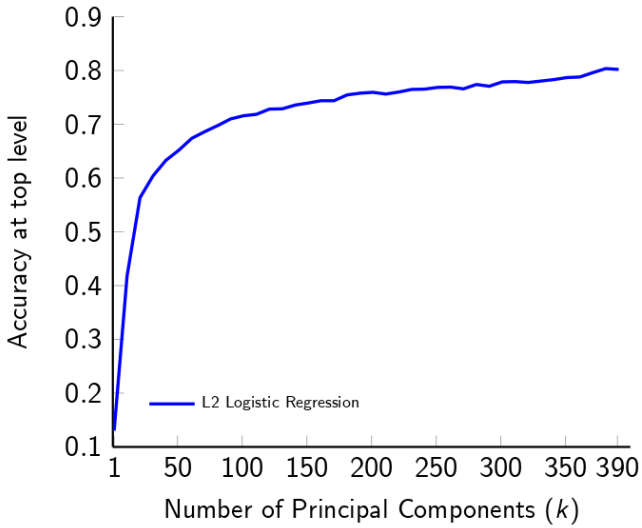
Стандартные подходы

- Отбор признаков с помощью χ^2 на верхних уровнях ведут к сильно разреженным векторам признаков для многих документов
- PCA требует спектрального разложения матрицы $X^T \cdot X$ $(p \times p)$, где X -тренировочная матрица
- Решение: EM PCA

Expectation Maximization PCA

- В EM PCA количество k главных компонент должно быть задано с самого начала
- $k << p$
- В EM PCA на каждом шаге одна из матриц имеет размерность равную k.
- p - исходное число признаков
- n - число документов

Уменьшение размерности на верхних уровнях



Number of Features	Accuracy
55,765 (100%)	0.82
27,882 (50%)	0.76
5,576 (10%)	0.56
557 (1%)	0.29

Table: Using χ^2 feature selection

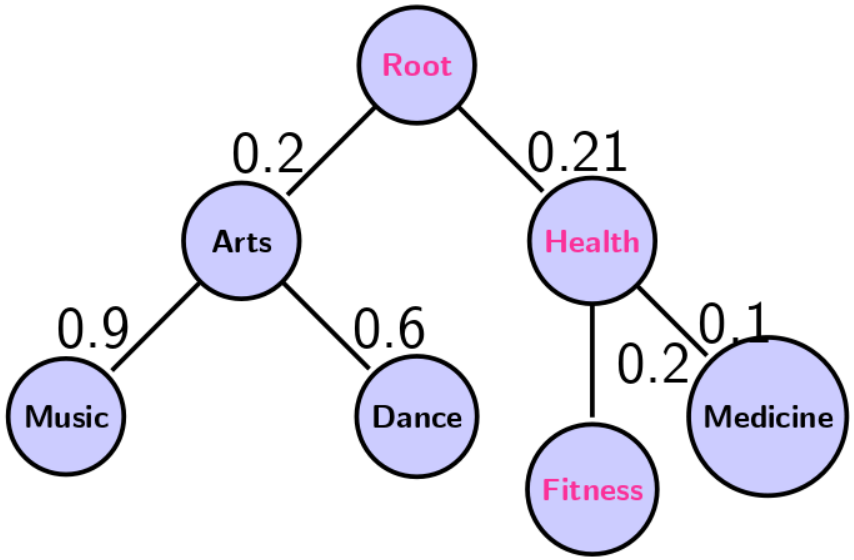
- LSHTC Dry-run Hierarchy
- Классификатор: L2 Regularized Logistic Regression

Эксперименты

Evaluation Measure	Cascade	PCA LR	Flat	PCA SVM
Accuracy	0.404*	0.385	**0.405***	0.402*
Macro F-measure	**0.278***	0.259	0.256*	0.274*
Macro Precision	**0.269**	0.249	0.254	0.264
Macro Recall	0.289	0.268	**0.302**	0.283
Tree Induced Error	**3.609**	3.845	3.874	3.616
Training Time	8.7 min	6.2 min	1017.6 min	127.5 min

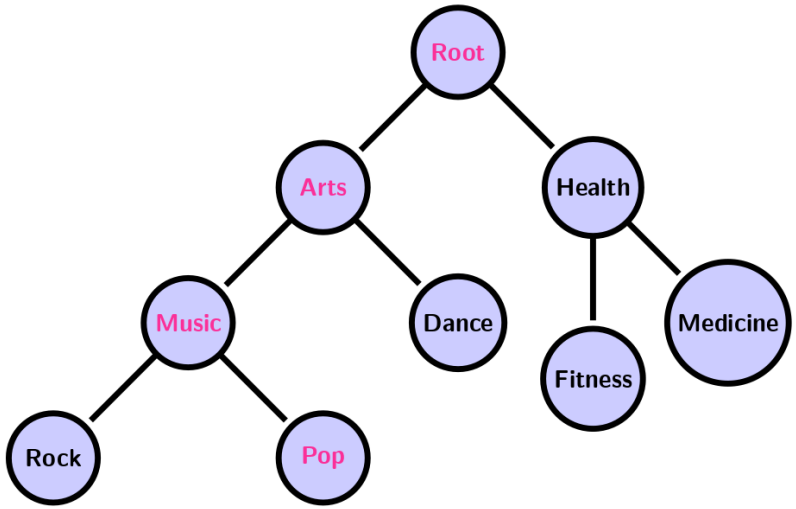
- LSHTC Large Hierarchy
- 381,581 признак
- 12,294 категорий
- 93,505 объекта в обучении
- $k = 490$

Каскад классификаторов с вероятностями



- Корректный класс: *Music*
- Предсказанный класс: *Fitness*

Каскад классификаторов с вероятностями



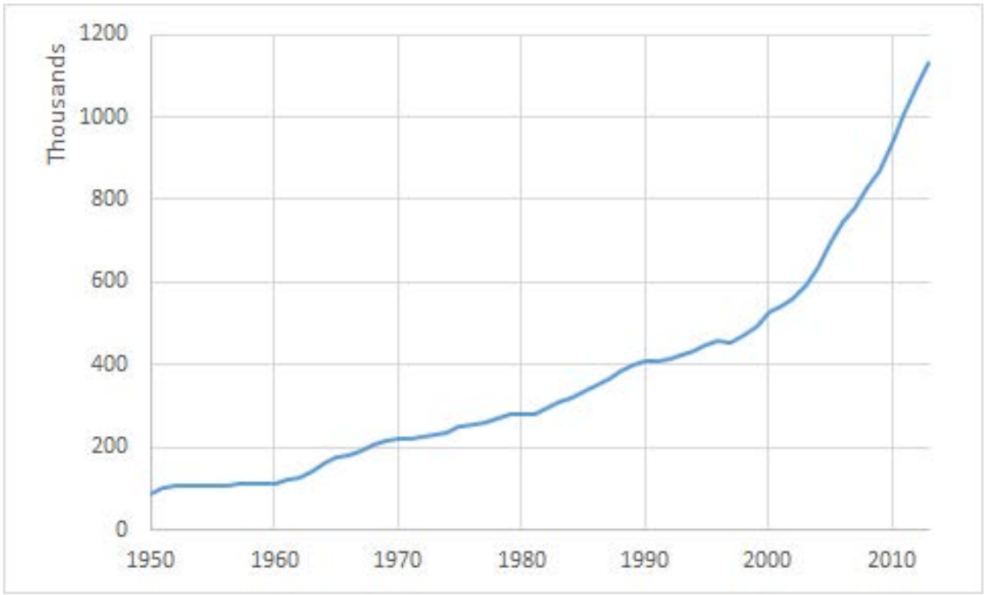
$$P(Pop|d) = P(Pop|Music, d) \cdot P(Music|Arts, d) \cdot P(Arts|Root, d)$$

Каскад классификаторов с вероятностями

Evaluation Measure	Flat	Cascade	P_{path}
Accuracy	0.405	0.404	**0.431**
Macro F-measure	0.256	0.278	**0.294**
Macro Precision	0.254	0.269	**0.287**
Macro Recall	**0.302**	0.289	**0.302**
Tree Induced Error	3.874	3.609	**3.437**

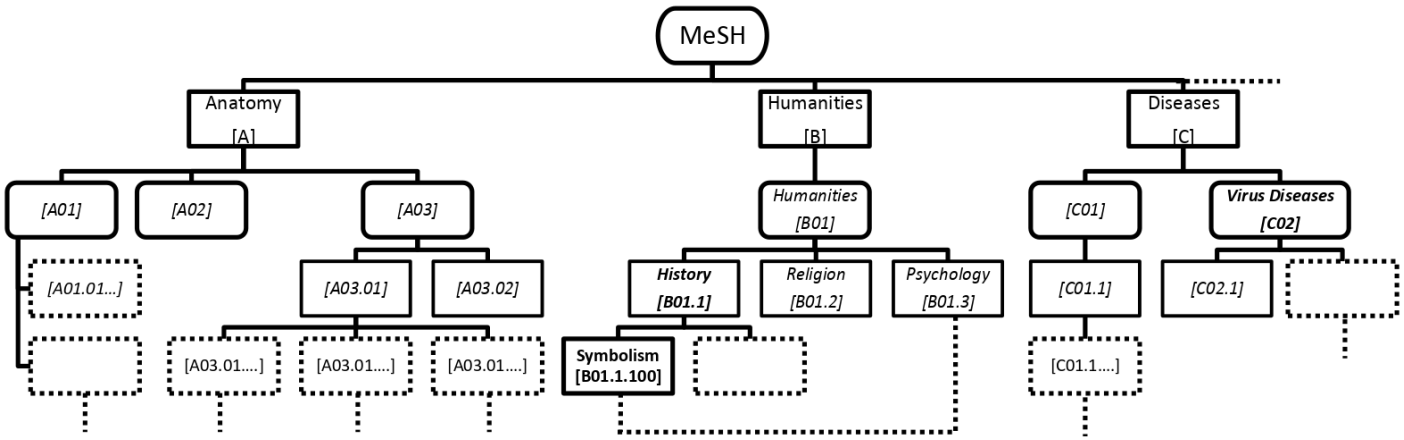
- LSHTC Large Hierarchy

MEDLINE



- MEDLINE - это база биомедицинских статей, созданная The United States National Library of Medicine.
- Сейчас более 24 миллионов статей
- Кураторам нужно вручную размечать статьи

Иерархия категорий



- Статьи могут принадлежать нескольким категориям
- Классификация возможна и во внутренние узлы

Классификация. Стандартные подходы

Проблемы:

- Bag of words: миллионы признаков
- Отбор признаков работает плохо, особенно на верхних уровнях
- Стемминг в биомедицинских текстах работает плохо

От мешка слов к векторам

Проблемы:

- В мешке слов каждое слово представляется разреженным вектором: $<0, 0, \ldots, 1, 0, 0, \ldots, 0>$
- Вложения слов: каждое слово представляется неразреженным вектором небольшой размерности (100-300)
- Преобразование осуществляется с помощью *word2vec* для 10,876,004 медицинских статей (BioASQ challenge)
- Получились вектора размерности 200 для 1,701,632 слов.

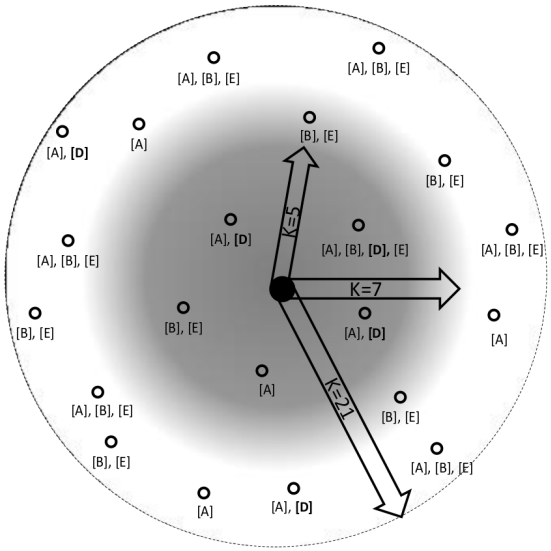
Нахождение ближайших слов

protein	**proteins**	**a-anchoring**	**pka-anchoring**
thyroid	**thyroidal**	**nonthyroid**	hyperfunctioning
associated	**correlated**	**related**	**correlates**
hormone	**gh**	**luetinizing**	**fshluteinizing**
human	murine	mouse	~~immortalized~~
used	**utilized**	**employed**	**applied**
genes	**gene**	**paralogs**	**operons**
treatment	**therapy**	**treatments**	**treating**
disease	**diseases**	disease-like	~~mmrn1rs6532197~~
gene	**genes**	**pseudogene**	**gene-encoding**
heart	**cardiac**	**chf**	**congestive**
role	**roles**	~~plays~~	~~play~~
affect	alter	modify	impair
dna	**dnas**	bisulfite-treated	polymerase-mediated
histone	**histones**	**h4k16**	**h4**
involved	**implicated**	**participates**	**regulating**
list	**lists**	**listing**	to-do
proteins	**protein**	**polypeptides**	**hsp70s**
known	~~yet~~	~~presently~~	well-known
patients	**outpatients**	**subjects**	whom
present	this	~~aimed~~	~~our~~
cancer	**cancers**	**crc**	~~caner~~
receptor	**receptors**	**hmc5**	5-nonyloxytryptamine
regulate	**modulate**	**regulates**	**orchestrate**
cell	**cells**	**cancer-cell**	**sw1710**
coding	**5-noncoding**	**5-untranslated**	**3-noncoding**
inhibitors	**inhibitor**	**small-molecule**	**atp-competing**
many	**several**	**some**	**numerous**
related	**linked**	**associated**	**relate**
cardiomyopathy	**cardiomyopathies**	**myocardiopathy**	**dcm**

Вектор для статьи

- Для получения вектора $\vec{t}$ для текста $t =< w_1, w_2, \ldots, w_n >$ из n последовательных слов вычисляется центроид:
- $\vec{t} = \frac{1}{n} \sum_{i=1}^{n} \vec{w}_i = \frac{\sum_{j=1}^{|V|} \vec{w}_j \cdot TF(w_j,t)}{\sum_{j=1}^{|V|} TF(w_j,t)}$
- $|V|$ - число уникальных слов в словаре
- $TF(w_j, t)$ - число вхождений w_j в текст t
- Вместе с обратной частотой документа:
- $\vec{t} = \frac{1}{n} \sum_{i=1}^{n} \vec{w}_i = \frac{\sum_{j=1}^{|V|} \vec{w}_j \cdot TF(w_j,t) \cdot IDF(w_j)}{\sum_{j=1}^{|V|} TF(w_j,t) \cdot IDF(w_j)}$

k-NN



- Эксперименты с евклидовой и с косинусной мерой сходства
- Класс связывается со статьей, если по крайней мере половина соседей принадлежит этому классу
- Величина K сильно зависит от «редких» классов

Результаты

	Dense TF-IDF Centroids	Paragraph Vectors	TF-IDF Bag-of-words
Micro F_1	**40.8%**	39.9%	38.1%
Micro Precision	62.2%	64.6%	61.9%
Micro Recall	30.4%	28.9%	27.6%
Macro F_1	17.1%	13.3%	**20.1%**
Macro Precision	51.7%	52.8%	59.7%
Macro Recall	15.2%	11.8%	17.9%
Accuracy	**26.3%**	25.3%	24.6%
LCA-F_1	**34.5%**	32.5%	31.8%
LCA Precision	53.7%	54.4%	53.1%
LCA Recall	27.5%	25.3%	24.8%
Classification Time	**50 min**	**50 min**	470 min
Preprocessing Time	**1 min**	9,960 min	**1 min**

- Везде использовался k-NN, $K = 5$, $p = 50\%$, косинусная мера сходства

sentence2vec

- Переводятся в вектор не только отдельно слова, но и целые параграфы (предложения, абзацы)
- 2 подхода:
  1. Вектор параграфа используется для предсказаний векторов слов
  2. Вектор параграфа и вектора слов из этого параграфа используются для предсказаний вектора следующего параграфа
- (Le and Mikolov, 2014): нужно использовать комбинацию этих двух подходов
- sentence2vec работает очень медленно

k-NN. Зависимость от числа соседей

	$K = 5$	$K = 7$	$K = 13$	$K = 21$
Micro F_1	**40.3%**	39.7%	38.3%	37.3%
Micro Precision	62.6%	66.0%	70.3%	**72.4%**
Micro Recall	**29.7%**	28.4%	26.4%	25.1%
Macro F_1	**15.9%**	14.3%	11.6%	9.9%
Macro Precision	52.2%	56.7%	62.8%	**66.4%**
Macro Recall	**13.9%**	12.2%	9.8%	8.2%
Accuracy	**25.8%**	25.3%	24.2%	23.4%
LCA-F_1	**33.6%**	32.3%	30.4%	29.0%
LCA Precision	53.9%	55.1%	56.6%	**57.2%**
LCA Recall	**26.6%**	24.8%	22.4%	20.9%

Итоги

- Две новые иерархические оценки качества предсказаний (MGIA, LCA)
- Эффективный метод уменьшения размерности в задаче иерахической классификации текстов
- Метод k ближайших соседей для классификации биомедицинских статей