0007 sheet | učo | points

Describe the SoundEx phonetic retrieval algorithm [2 points]. Give an example of two similarly-sounding words with the same SoundEx code [1 point]. Explain the weak point of SoundEx [1 point], and give an example of two similarly-sounding words with different SoundEx codes [1 point].

Generally speaking, the SoundEx algorithm maps similarly-sounding words together by removing vowels and silent consonants (W and H), and clustering consonants with similar English spelling together.

sword and short have code S630.

The major weak point of the SoundEx algorithm is the dependency on the first letter of a word that is retained.

Two different SoundEx codes for similarly-sounding words are fog (F200) and thug (T200).

Consider the following collection of four documents d_i:

- d_1: BREAKTHROUGH DRUG FOR HIV
- d_2: NEW HIV DRUG
- d_3: NEW APPROACH FOR TREATMENT OF HIV
- d_4: NEW HOPES FOR HIV PATIENTS

Produce a list of (term, document ID) tuples [1 point], sort this list in lexicographical order [1 point], and use the sorted list to construct an inverted index [1 point]. Write down each step. Describe how you would produce this index using the MapReduce distributed framework [2 points].

(breakthrough, 1), (drug, 1), (for, 1), (HIV, 1),
(new, 2), (HIV, 2), (drug, 2),
(new, 3), (approach, 3), (for, 3), (treatment, 3), (of, 3), (HIV, 3),
(new, 4), (hopes, 4), (for, 4), (HIV, 4), (patients, 4)

(approach, 3), (breakthrough, 1), (drug, 1), (drug, 2), (for, 1),
(for, 3), (for, 4), (HIV, 1), (HIV, 2), (HIV, 3), (HIV, 4),
(hopes, 4), (new, 2), (new, 3), (new, 4), (of, 3), (patients, 4),
(treatment, 3)

approach → 3
breakthrough → 1
drug → 1 → 2
for → 1 → 3 → 4
HIV → 1 → 2 → 3 → 4
hopes → 4
new → 2 → 3 → 4
of → 3
patients → 4
treatment → 3

Each parser would process a limited number of documents and produce a sorted (term, document ID) list.
Each inverter would take all sublists in a certain alphabetical range of terms and produce postings for that alphabetical range.

0007 sheet 3 učo points

Define the two assumptions the *Naive Bayes* classifier makes [2 points]. Explain the advantage of computing a product of probability estimates as a sum in the logarithmic space [1 point]. Given an observation $\mathbf{x}$, and the classes c_1, and c_2, is the knowledge of $P(\mathbf{x} \mid c_1)P(c_1) > P(\mathbf{x} \mid c_2)P(c_2)$ sufficient to decide whether $P(c_1 \mid \mathbf{x}) > P(c_2 \mid \mathbf{x})$? Why or why not? [2 points]

Given the following list of observations, use the Naive Bayes classifier to decide whether to play golf when it is sunny, hot, windy, and the humidity is normal. [5 points]

Outlook	Temperature	Humidity	Windy	Play golf
Sunny	Mild	High	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	True	No
Sunny	Mild	Normal	False	Yes
Rainy	Cool	Normal	False	Yes
Overcast	Hot	High	False	Yes
Rainy	Hot	High	False	No
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Cool	Normal	True	No
Rainy	Hot	High	True	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	High	False	No

Two assumptions: conditional and positional independence.

Advantage: numerical stability of multiplying small real numbers.

By Bayes' theorem, $P(c_i \mid \vec{x}) = \frac{P(\vec{x} \mid c_i) \cdot P(c_i)}{P(\vec{x})}$ for some $i \in \{1,2\}$.

Therefore, $P(\vec{x} \mid c_1) \cdot P(c_1) > P(\vec{x} \mid c_2) \cdot P(c_2)$ implies $P(c_1 \mid \vec{x}) > P(c_2 \mid \vec{x})$, and vice versa.

$P(\text{YES} \mid \text{Sunny}, \text{Hot}, \text{Normal}, \text{True}) = P(\text{Sunny} \mid \text{YES}) \cdot P(\text{Hot} \mid \text{YES}) \cdot P(\text{Normal} \mid \text{YES}) \cdot P(\text{True} \mid \text{YES}) \cdot P(\text{YES}) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} = \frac{6}{81.7}$

$P(\text{NO} \mid \text{Sunny}, \text{Hot}, \text{Normal}, \text{True}) = P(\text{Sunny} \mid \text{NO}) \cdot P(\text{Hot} \mid \text{NO}) \cdot P(\text{Normal} \mid \text{NO}) \cdot P(\text{True} \mid \text{NO}) \cdot P(\text{NO}) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} = \frac{6}{125 \cdot 7}$

$P(\text{YES} \mid \text{Sunny}, \text{Hot}, \text{Normal}, \text{True}) > P(\text{NO} \mid \text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$;

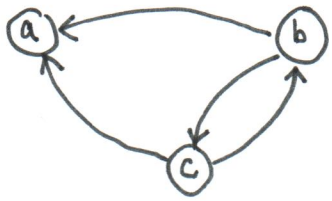
Therefore, play golf.

0007 sheet 4 učo points

Given a directed graph G that represents three Web pages $V(G) = \{a, b, c\}$, and the links $E(G) = \{(b,a), (c,a), (c,b), (b,c)\}$ between these three pages, draw G [1 point] and produce *the adjacency matrix* (also known as *the link matrix*) **A** [1 point], and *the Markov transition matrix* **P** [2 points].

Describe the intuition behind *the PageRank algorithm* [1 point]. Compute the *PageRank* of the pages a, b, and c using a single iteration of the PageRank algorithm [2 points].

Describe what we mean, when we call a page a *hub*, or an *authority* [1 point]. Compute the *hub*, and *authority scores* of the pages a, b, and c [2 points].



Graph G:

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$P = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \circ \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 \end{bmatrix} \cdot (1-\alpha) + \frac{\alpha}{3}$$

where $\circ$ is the Hadamard product.

The PageRank algorithm computes the probability that a hypothetical random surfer will end up at a given web page.

$\vec{x}_0 = (1\ 0\ 0)$ $\quad \vec{x}_1 = \vec{x}_0 \cdot P = [1\ 0\ 0] \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ \cdots & & \end{bmatrix} = [\frac{1}{3}\ \frac{1}{3}\ \frac{1}{3}]$

A hub is a web page pointing to many authorities.

An authority is a web page that many hubs point to.

$\vec{h}_0 = [1\ 1\ 1]^T$

$\vec{a}_0 = [1\ 1\ 1]^T$

$$A \cdot A^T = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

$$A^T \cdot A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$\vec{h}_1 = A \cdot A^T \cdot \vec{h}_0 = [0\ 3\ 3]^T \approx [0\ 1\ 1]^T$

$\vec{a}_1 = A^T \cdot A \cdot \vec{a}_0 = [4\ 2\ 2]^T \approx [1\ \frac{1}{2}\ \frac{1}{2}]^T$