0007 sheet 1 učo 409729 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 different-sounding words with the same SoundEx codes [1 point].

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

Jword and short have code S63Ø.

The major weak point of SoundEx is the fact that the initial letter of a word is retained.

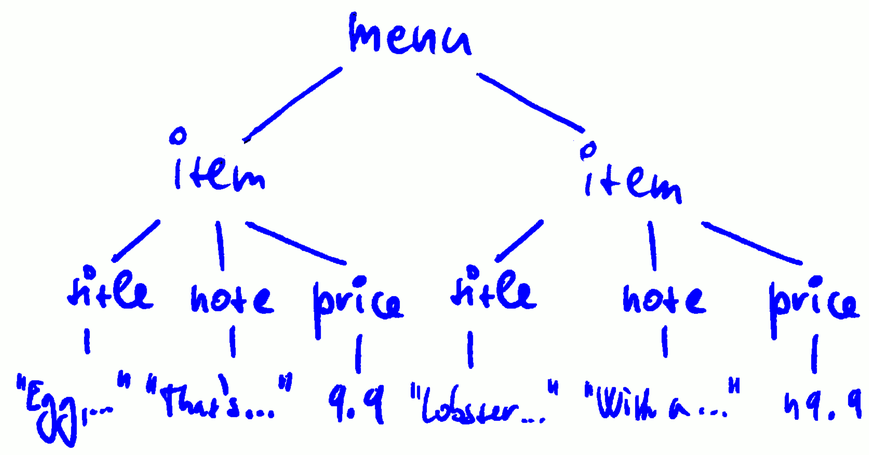
Two different-sounding words with the same SoundEx codes are Stefan (S315), and Steven (S315).

You may also write on the other side of the sheet, but it won't be scanned.

0007 sheet 2 učo 409729 points

Consider the following XML document:

```
<?xml version="1.0" encoding="UTF-8"?>
<menu>
    <item>
        <title>Egg, sausage, and spam</title>
        <note>That's not got much spam in it</note>
        <price unit="GBP">9.9</price>
    </item>
    <item>
        <title>Lobster Thermidor</title>
        <note>With a fried egg on top, and spam</note>
        <price unit="GBP">49.9</price>
    </item>
</menu>
```



Draw the XML document as a graph [1 point], and count all *structural terms* in the document [2 points]. Compute the *structural similarity* between the structural term `item/title#"That's not got much spam in it"`, and the queries `//menu//price#30.9`, and `//item/title#"Lobster Thermidor"` [2 points].

Structural terms: #"Egg,...", title#"Egg,...", item/title#"Egg,...", menu/item/title#"Egg,...", #"that's...", note#"that's...", item/note#"that's...", menu/item/note#"that's...", #9.9, price#9.9, item/price#9.9, menu/item/price#9.9, #"Lobster...", title#"Lobster...", item/title#"Lobster...", menu/item/title#"Lobster...", #"With a...", note#"With a...", item/note#"With a...", menu/item/note#"With a...", #49.9, price#49.9, item/price#49.9, menu/item/price#49.9

6·4 = 24 structural terms.

//menu//price does not expand to item/title, the structural similarity is zero. //item/title is equivalent to item/title, so the structural similarity is $\frac{1+1}{1+1} = 1$.

You may also write on the other side of the sheet, but it won't be scanned.

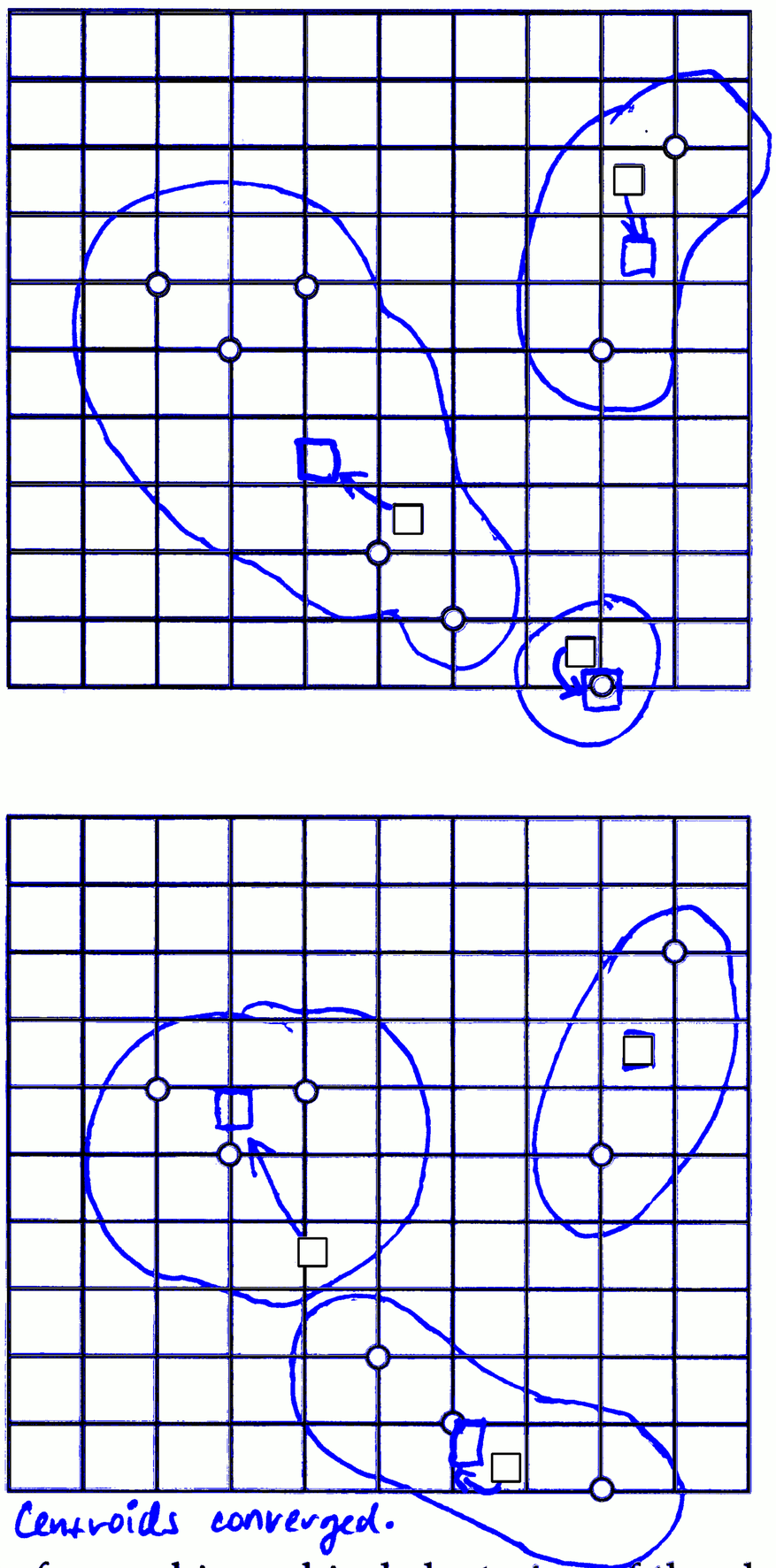
0007 *sheet* 3 *učo* 409729 *points*

Explain the following aspects of the K-means flat clustering algorithm [2 points]:

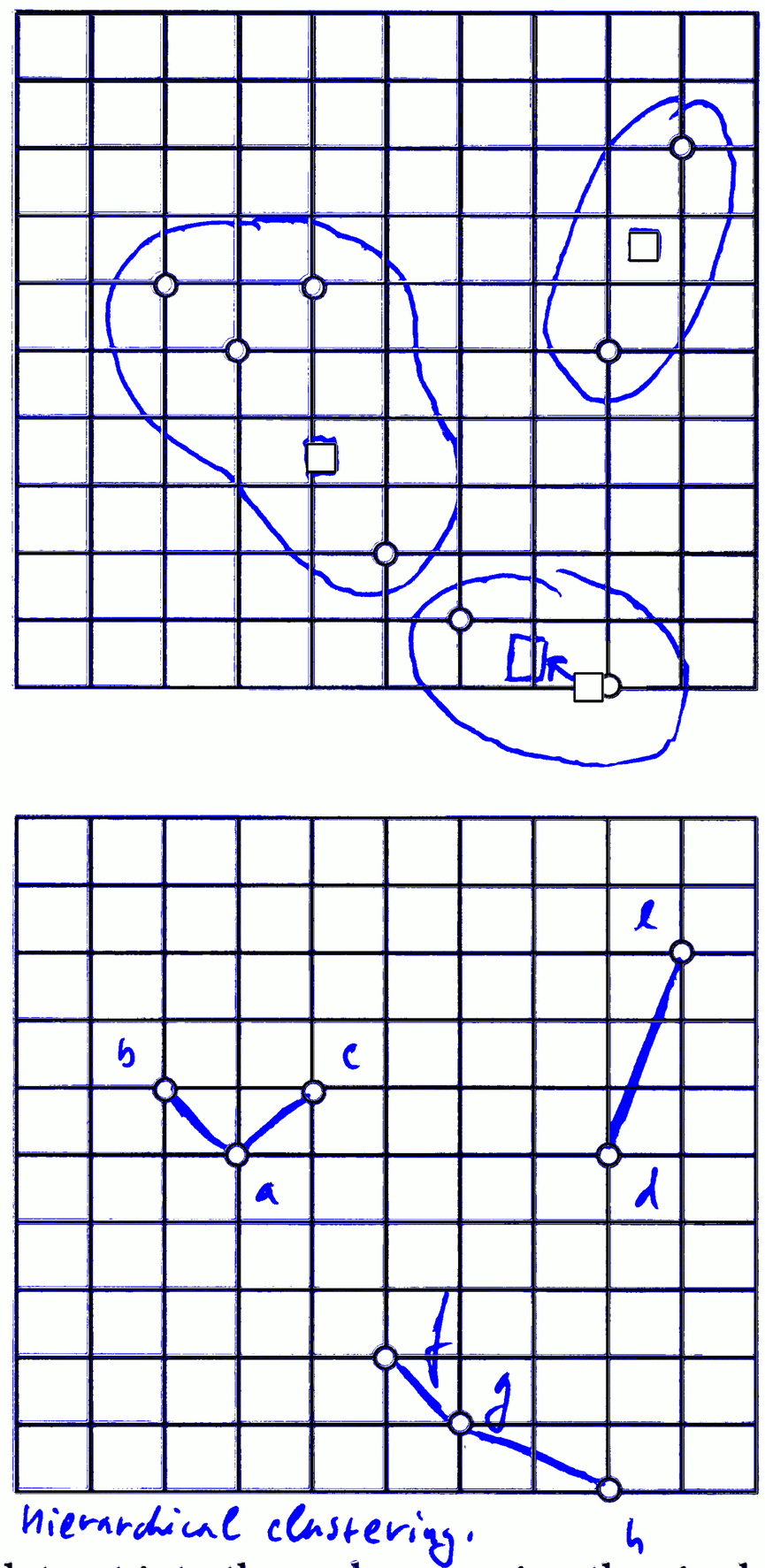
1. What do we need to know about our dataset before using the algorithm?
2. What is the input and the output of the algorithm?
3. What are the two steps that take place in every epoch?
4. How do we decide in which epoch to stop the algorithm?

1. We need to know the number of classes K and initial mean estimates (seeds). 2. The input are unclassified points and K seeds. 3. Reassigning points, recomputing centroids. 4. Centroids converged.

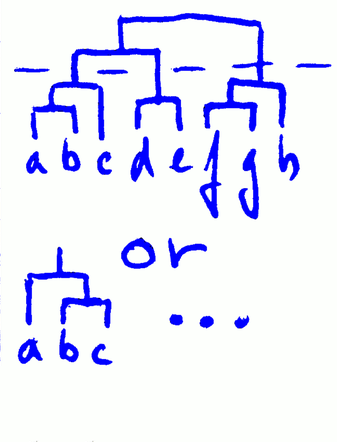
Given the points ○, and the seeds □, run the K-means algorithm for three epochs. Draw the state of the algorithm at the beginning and after every epoch; no computation should be necessary. What is the output of the algorithm? [2 points]



Centroids converged.



Hierarchical clustering.



Perform a hierarchical clustering of the above dataset into three classes using the single-link hierarchical agglomerative clustering algorithm, and draw the resulting *dendrogram*. [1 point] Is the output the same as the output of the K-means flat hierarchical clustering algorithm above? [1 point]

Yes, it is.

You may also write on the other side of the sheet, but it won't be scanned.

0007 sheet 4 učo 409729 points

State the *Zipf's law* [1 point] and the *Heaps' law* [1 point] in its complete variant [1 point]. Let **C** denote the *term-document matrix* of a document collection that contains M unique terms. According to the Heaps' law, what is the number of documents N in the collection in relation to M? [3 points]

Zipf's law: $cf_i = c/i$, where cf_i is the collection frequency of the ith most frequent term, and c is a constant.

Heaps' law: $M = kT^b$, where M is the vocabulary size, T is the collection size in tokens, and $30 \leq k \leq 100$ and $b \approx 0,5$ are parameters.

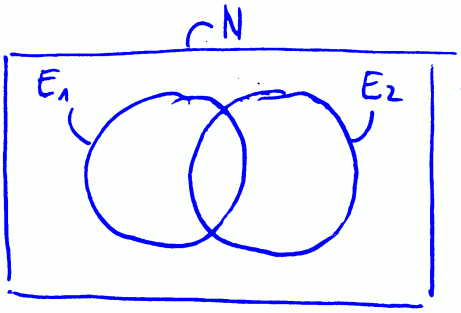
The size of **C** is $M \times N < T$, where N is the number of documents in a collection. According to the Heaps' law, $T = \sqrt[b]{M/k} \approx M^2$. Therefore $M \times N < M^2 \Rightarrow N < M$.

You may also write on the other side of the sheet, but it won't be scanned.

0007 sheet 5 učo 4097129 points

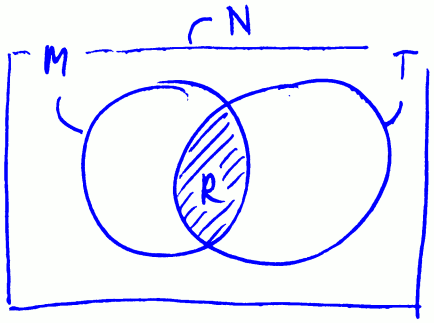
You are the maintainer of a text retrieval system. Let E_1 denote the complete set of documents in the index of your system and let E_2 denote the complete set of documents in the index of a competing system. Suppose the indices of both systems are independent uniform random samples without replacement from the World Wide Web N. The size of E_1 is $|E_1| = 130$ trillion $(130 \cdot 10^{12})$ documents. You take a uniform random subsample of documents without replacement from E_1 and you submit each document to the competing system. This gives you an unbiased estimate $x = 0.2$ of the conditional probability $P(d \in E_2 \mid d \in E_1), d \in N$. You repeat the same procedure with E_2, obtaining an unbiased estimate $y = 0.4$ of the conditional probability $P(d \in E_1 \mid d \in E_2), d \in N$. Assume the estimates x, y are the true probabilities. What is the size $|E_2|$ of the competing system's index? [4 points]

The grey parrot, native to equatorial Africa, is categorized as an endangered species by the International Union for Conservation of Nature (IUCN). Suppose you take a uniform random sample M without replacement of size $|M| = 8\,000$ from the grey parrot population N and mark the sampled animals. After returning the marked animals back into the population, you take a second independent uniform random sample T without replacement of the same size $|T| = 8\,000$ from the population. The number of marked animals $R = M \cap T$ in the second sample is $|R| = 10$. What is the size $|N|$ of the population of the grey parrot? [4 points]



$\forall d \in N: P(d \in E_2 \mid d \in E_1) = 0{,}2,$
$P(d \in E_1 \mid d \in E_2) = 0{,}4$
$P(d \in E_1) = \frac{130 \cdot 10^{12}}{|N|}, \quad P(d \in E_2) = \frac{|E_2|}{|N|}.$

$0{,}2 = \frac{0{,}4 \cdot \frac{|E_2|}{|N|}}{\frac{130 \cdot 10^{12}}{|N|}} \Leftrightarrow |E_2| = \frac{0{,}2 \cdot 130 \cdot 10^{12}}{0{,}4} = 65 \cdot 10^{12}.$



$\forall x \in N: P(x \in M) = \frac{|M|}{|N|}, \quad P(x \in R) = \frac{|R|}{|N|}$
$P(x \in T) = \frac{|T|}{|N|}.$

$\forall x, y \in N: P(x \in R) = P(x \in M, y \in T)$
$= P(x \in M) \cdot P(y \in T) = \frac{|M||T|}{|N|^2}.$

$\frac{|R|}{|N|} = \frac{|M||T|}{|N|^2} \Leftrightarrow |N| = \frac{|M||T|}{|R|} = \frac{8000^2}{10} = 6{,}4 \cdot 10^6.$

Write your solution only on this side of the sheet!