INTRODUCTION TO DATA SCIENCE

This lecture is based on course by E. Fox and C. Guestrin, Univ of Washington

12/01/2022

WFAiS UJ, Informatyka Stosowana I stopień studiów

Retrieving documents of interest

2

- Currently reading article you like
- Goal: Want to find similar article





Retrieving documents of interest

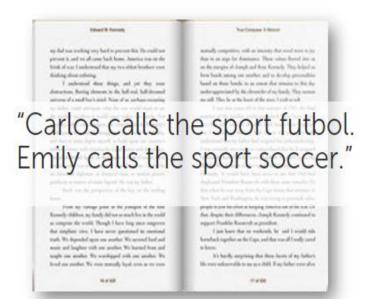
Challenges

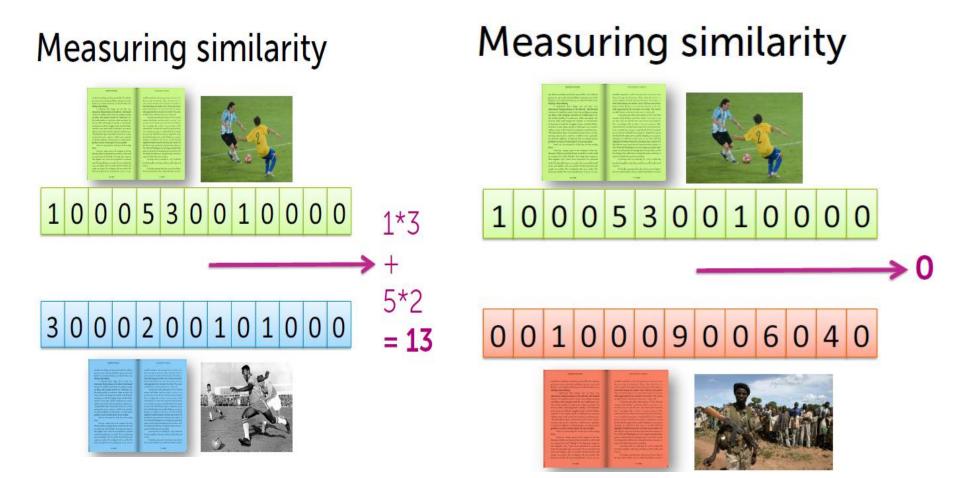
- How do we measure similarity?
- How do we search over articles?



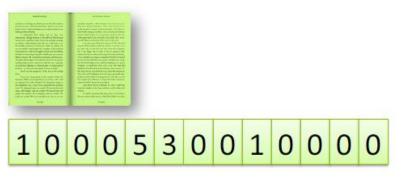
- Bag of words model
 - Ignore order of words
 - Count # of instances of each word in vocabulary

.... 6 2 D

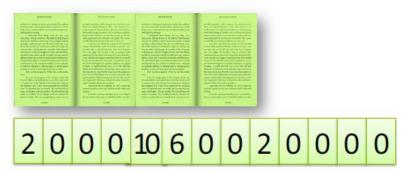


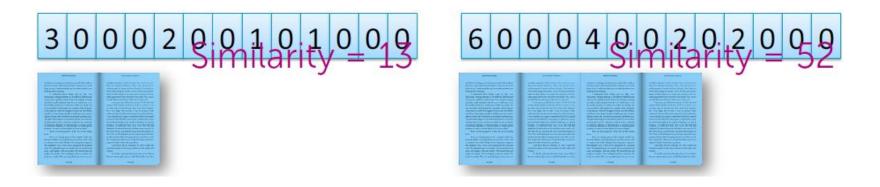


Issues with word counts - Doc length

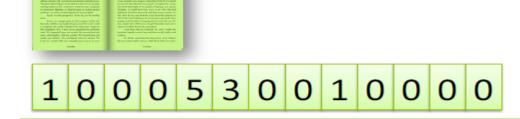


6





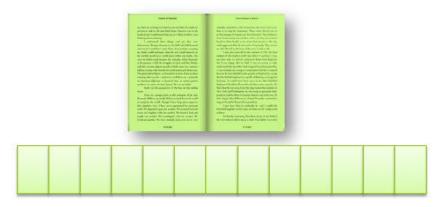
Solution = normalize



$$\sqrt{(1^2 + 5^2 + 3^2 + 1^2)}$$

Prioritizing important words

Issues with word counts – Rare words



Common words in doc: "the", "player", "field", "goal" Dominate rare words like: "futbol", "Messi"

Prioritizing important words

Document frequency

- What characterizes a rare word?
 - Appears infrequently in the corpus

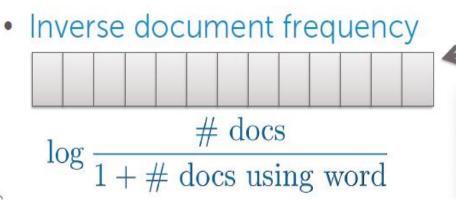
Prioritizing important words

Important words

- Do we want only rare words to dominate???
- What characterizes an important word?
 - Appears frequently in document (common locally)
 - Appears rarely in corpus (rare globally)
- Trade off between local frequency and global rarity

TF-IDF document representation

- Term frequency inverse document frequency (tf-idf)
- Term frequency

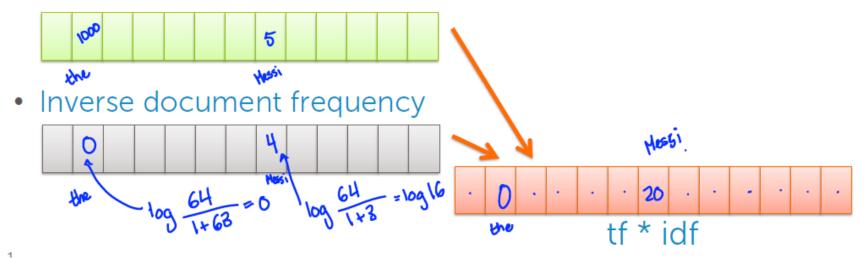




TF-IDF document representation

TF-IDF document representation

- Term frequency inverse document frequency (tf-idf)
- Term frequency



Retrieving similar documents

anteriorente attentionen

Nearest neighbor search

- Query article:
- Corpus:

transfer a Lashani da siya dashinin na Antiq kalendari Yoshini karang, si'ar dar sar	 Interface of Anial La (Kar Kaar), No Ellis Li Interface and an estimate after the improvements interface depicted (in Control of Land of Land). 		
denotes the policy of the life of high sector of a sector of policy of the life of the sector of the life of the l	well finds our hand of sectors and		
According specific to contract, while the baseline of the base	Construction of the Association		
a degree of the second state of the second state. and has a face for each of the second state of the second state to be a second state of the second state of the second	see 1 con logge data on the 1 con an array of the measurements of the log operation are seen in the logge meeting of the provided for the measure		
- Watchington and the second second	Strate Strate Photo State State of State States		
en handwar skillen in fennet han en selle samt sellerer ei som di som igget i förstag allen Red var föransarberd äft för en ett reden	Station, Yound Weit, Son and in an Andread Intelligence Systematic Learning Systems on American Systems (Sciences), 5 No. American Systems, Space Systems, Strangence, Stra		
the second second second second second	And finds and Mindolphic & new York, provide the analysis of the description of the second s		
Annual State of the second sec	An Anna San Maran (Andreas) and and an Instance for the second se		
with Wings and again and Wings and Wings and the second se	And a state of the		
And the setting is a straight for the section	And the state of t		
THE OWN STATISTICS.	and according on a second	A ROUTE AND A ROUT	AND REPORTED AND AND AND AND AND AND AND AND AND AN
ALL STREET, ST	THE REAL PROPERTY	HILL MANY CONTRACTOR	A ARACE DIS
THEFTER OF CHERTER	THE REPORT AND INCOME.	BEPTER PROPERTY OF	THE STREET NO.
		이 비행 위한 비행	
CHARLES CHARLES	A MARKAGAMARA A	A MARAGENIAR	ALCONY DECHICLE A RANK
NAMES AND A STREET AND A DECK		ACCESSION NO.	THE SHE AT
Internal and the set of the strength of a second state of the state of the			
THE REPORT OF THE REPORT OF THE PROPERTY OF TH			
Himselfer annual and	inere and entered the	STREET, STREET, STORE	Statistic strategies and a
	在 各国的新闻的新闻	FIRST IN THE REAL	ST IN STATIST
A STATE OF STATE OF STATE	and a post to and the first of the	A MANAGARANA PARAMANA	The local division of the local
A Part of the second		No. of Concession, Name	
	and the second se		Statement of the statem

- Specify: Distance metric
- Output: Set of most similar articles



Retrieving similar documents

1 – Nearest neighbor

Input: Query article



- Output: Most similar article
- Algorithm:
 - Search over each article 📃 in corpus
 - Compute **s = similarity(**
 - If s > Best_s, record and set Best_s = s





Retrieving similar documents

k – Nearest neighbor

- Input: Query article
- Output: List of k similar articles



Structure documents by topics

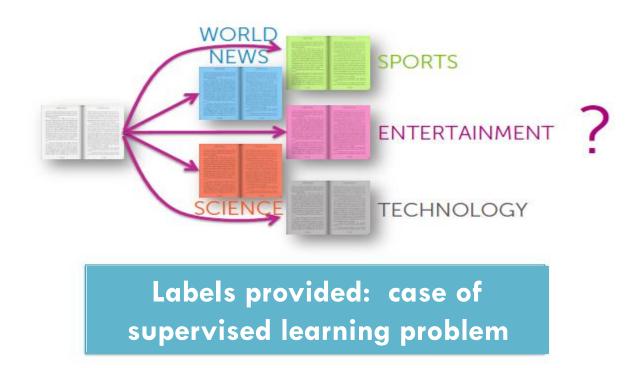
What if some of the labels are known?

• Training set of labeled docs



Structure documents by topics

Multiclass classification problem



Clustering

- No labels provided
- Want to uncover cluster structure
- Input: docs as vectors
- Output: cluster labels

fluster

No labels provided unsupervised learning

12/01/2022

word

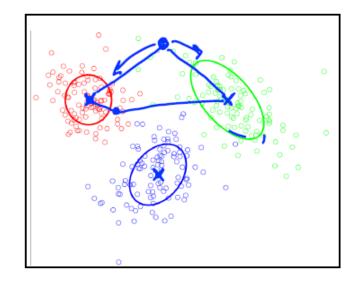
2

ℑ

Clustering

What defines a cluster?

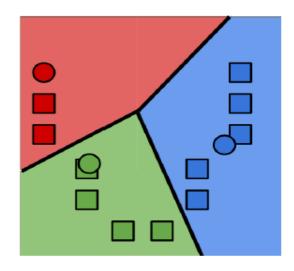
- Cluster defined by center & shape/spread
- Assign observation (doc) to cluster (topic label)
 - Score under cluster is higher than others
 - Often, just more similar to assigned cluster center than other cluster centers



Clustering

k-means algorithm

- 0. Initialize cluster centers
- 1. Assign observations to closest cluster center
- 2. Revise cluster centers as mean of assigned observations
- 3. Repeat 1.+2. until convergence

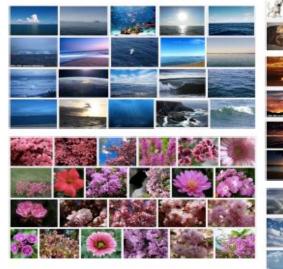


Clustering images

- For search, group as:
 - Ocean
 - Pink flower
 - Dog

- ...

- Sunset
- Clouds





Products on Amazon

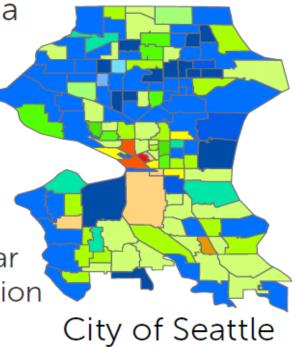
 Discover product categories from purchase histories



Or discovering groups of users

Discovering similar neighborhoods

- Task 1: Estimate price at a small regional level
- Challenge:
 - Only a few (or no!) sales in each region per month
- Solution:
 - Cluster regions with similar trends and share information within a cluster



Discovering similar neighborhoods

- Task 2: Forecast violent crimes to better task police
- Again, cluster regions and share information!
- Leads to improved predictions compared to examining each region independently

s the second sec

Washington, DC

We discussed how to ...

- Describe ways to represent a document (e.g., raw word counts, tf-idf,...)
- Measure the similarity between two documents
- Discuss issues related to using raw word counts
 - Normalize counts to adjust for document length
 - Emphasize important words using tf-idf
- Implement a nearest neighbor search for document retrieval
- Describe the input (unlabeled observations) and output (labels) of a clustering algorithm
- Determine whether a task is supervised or unsupervised
- Cluster documents using k-means (algorithmic details to come...)
- Describe other applications of clustering