CSE 158 – Lecture 17
Web Mining and Recommender Systems

More temporal dynamics
Temporal models

This week we’ll look back on some of the topics already covered in this class, and see how they can be adapted to make use of **temporal** information

1. **Regression** – sliding windows and autoregression
2. **Classification** – dynamic time-warping
3. **Dimensionality reduction** – ?
4. **Recommender systems** – some results from Koren

Today:

1. **Text mining** – “Topics over Time”
2. **Social networks** – densification over time
Monday: Time-series regression

Also useful to plot data:

BeerAdvocate, ratings over time

Scatterplot

BeerAdvocate, ratings over time

Sliding window (K=10000)

long-term trends

seasonal effects

Code on:
http://jmcauley.ucsd.edu/cse258/code/week10.py
Monday: Time-series classification

As you recall...
The longest-common subsequence algorithm is a standard dynamic programming problem

<table>
<thead>
<tr>
<th></th>
<th>-</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

1st sequence

2nd sequence

= optimal move is to delete from 1st sequence
= optimal move is to delete from 2nd sequence
= either deletion is equally optimal
= optimal move is a match
Monday: Temporal recommendation

To build a reliable system (and to win the Netflix prize!) we need to account for **temporal dynamics**:

(Netflix changed their interface)

(People tend to give higher ratings to older movies)

Figure from Koren: “Collaborative Filtering with Temporal Dynamics” (KDD 2009)
yeast and minimal red body thick light a Flavor sugar strong quad. grape over is molasses lace the low and caramel fruit Minimal start and toffee. dark plum, dark brown Actually, alcohol Dark oak, nice vanilla, has brown of a with presence. light carbonation. bready from retention. with finish. with and this and plum and head, fruit, low a Excellent raisin aroma Medium tan

Bags-of-words

What we would like:

Topic models

Sentiment analysis
8. Social networks

Hubs & authorities

Power laws

Small-world phenomena

Strong & weak ties
9. Advertising

Matching problems

AdWords

Bandit algorithms
Temporal dynamics of text
Bag-of-Words representations of text:

\[ F_{\text{text}} = [150, 0, 0, 0, 0, 0, 0, \ldots, 0] \]

- a
- aardvark
- zoetrope
In week 5/7, we tried to develop low-dimensional representations of documents:

**What we would like:**

87 of 102 people found the following review helpful

⭐⭐⭐⭐⭐ **You keep what you kill**, December 27, 2004

By Schtinky "Schtinky" (Washington State) - See all my reviews

This review is from: *The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)*

Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from 'Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to 'Pitch Black' fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of "The Chronicles of Riddick")
We saw how **LDA** can be used to describe documents in terms of **topics**

- Each document has a **topic vector** (a stochastic vector describing the fraction of words that discuss each topic)
- Each topic has a **word vector** (a stochastic vector describing how often a particular word is used in that topic)
Topics and documents are both described using stochastic vectors:

Each document has a topic distribution which is a mixture over the topics it discusses

\[ \theta_d \in \Delta^K \] i.e., \( \forall d \sum_k \theta_{d,k} = 1 \)

Each topic has a word distribution which is a mixture over the words it discusses

\[ \phi_k \in \Delta^D \] i.e., \( \forall k \sum_w \phi_{k,w} = 1 \)
Topics over Time (Wang & McCallum, 2006) is an approach to incorporate temporal information into topic models

e.g.
• The topics discussed in conference proceedings progressed from neural networks, towards SVMs and structured prediction (and back to neural networks)
• The topics used in political discourse now cover science and technology more than they did in the 1700s
• With in an institution, e-mails will discuss different topics (e.g. recruiting, conference deadlines) at different times of the year
Latent Dirichlet Allocation

**Topics over Time** (Wang & McCallum, 2006) is an approach to incorporate temporal information into topic models.

The ToT model is similar to LDA with one addition:

1. For each topic $K$, draw a word vector $\phi_k$ from Dir.$(\beta)$
2. For each document $d$, draw a topic vector $\theta_d$ from Dir.$(\alpha)$
3. For each word position $i$:
   1. draw a topic $z_{di}$ from multinomial $\theta_d$
   2. draw a word $w_{di}$ from multinomial $\phi_{z_{di}}$
   3. **draw a timestamp** $t_{di}$ **from Beta**$(\psi_{z_{di}})$
Topics over Time (Wang & McCallum, 2006) is an approach to incorporate temporal information into topic models:

3.3. draw a timestamp $t_{di}$ from Beta($\psi_{z_{di}}$)

- There is now one Beta distribution per topic
- Inference is still done by Gibbs sampling, with an outer loop to update the Beta distribution parameters

Beta distributions are a flexible family of distributions that can capture several types of behavior – e.g. gradual increase, gradual decline, or temporary “bursts”

$$p.d.f.: \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$
**Results:**
Political addresses – the model seems to capture realistic “bursty” and gradually emerging topics

<table>
<thead>
<tr>
<th>Mexican War</th>
<th>Panama Canal</th>
<th>Cold War</th>
<th>Modern Tech</th>
</tr>
</thead>
<tbody>
<tr>
<td>states 0.02032</td>
<td>government 0.02928</td>
<td>world 0.01875</td>
<td>energy 0.03902</td>
</tr>
<tr>
<td>mexico 0.01832</td>
<td>united 0.02132</td>
<td>states 0.01717</td>
<td>national 0.01534</td>
</tr>
<tr>
<td>government 0.01670</td>
<td>states 0.02067</td>
<td>security 0.01710</td>
<td>development 0.01448</td>
</tr>
<tr>
<td>united 0.01521</td>
<td>islands 0.01167</td>
<td>soviet 0.01664</td>
<td>space 0.01436</td>
</tr>
<tr>
<td>war 0.01059</td>
<td>canal 0.01014</td>
<td>united 0.01491</td>
<td>science 0.01227</td>
</tr>
<tr>
<td>congress 0.00951</td>
<td>american 0.00872</td>
<td>nuclear 0.01454</td>
<td>technology 0.01227</td>
</tr>
<tr>
<td>country 0.00906</td>
<td>cuba 0.00834</td>
<td>peace 0.01408</td>
<td>oil 0.01178</td>
</tr>
<tr>
<td>texas 0.00852</td>
<td>made 0.00747</td>
<td>nations 0.01069</td>
<td>make 0.00994</td>
</tr>
<tr>
<td>made 0.00727</td>
<td>general 0.00731</td>
<td>international 0.01024</td>
<td>effort 0.00969</td>
</tr>
<tr>
<td>great 0.00611</td>
<td>war 0.00660</td>
<td>america 0.00987</td>
<td>administration 0.00957</td>
</tr>
</tbody>
</table>
**Latent Dirichlet Allocation**

**Results:**

e-mails & conference proceedings

<table>
<thead>
<tr>
<th>Faculty Recruiting</th>
<th>ART Paper</th>
<th>MALLET</th>
<th>CVS Operations</th>
<th>Recurrent NN</th>
<th>Game Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs 0.03572</td>
<td>xuerni 0.02113</td>
<td>code 0.05668</td>
<td>check 0.04473</td>
<td>state 0.05963</td>
<td>game 0.02850</td>
</tr>
<tr>
<td>april 0.02724</td>
<td>data 0.01814</td>
<td>files 0.04212</td>
<td>page 0.04070</td>
<td>recurrent 0.03765</td>
<td>strategy 0.02378</td>
</tr>
<tr>
<td>faculty 0.02341</td>
<td>word 0.01601</td>
<td>mallet 0.04073</td>
<td>version 0.03828</td>
<td>sequence 0.03616</td>
<td>play 0.01490</td>
</tr>
<tr>
<td>david 0.02012</td>
<td>research 0.01408</td>
<td>java 0.03085</td>
<td>cvs 0.03587</td>
<td>sequences 0.02462</td>
<td>games 0.01473</td>
</tr>
<tr>
<td>lunch 0.01766</td>
<td>topic 0.01366</td>
<td>file 0.02947</td>
<td>add 0.03083</td>
<td>time 0.02402</td>
<td>player 0.01451</td>
</tr>
<tr>
<td>schedule 0.01656</td>
<td>model 0.01238</td>
<td>al 0.02479</td>
<td>update 0.02539</td>
<td>states 0.02057</td>
<td>agents 0.01346</td>
</tr>
<tr>
<td>candidate 0.01560</td>
<td>andres 0.01238</td>
<td>directory 0.02080</td>
<td>latest 0.02519</td>
<td>transition 0.01300</td>
<td>expert 0.01281</td>
</tr>
<tr>
<td>talk 0.01355</td>
<td>sample 0.01152</td>
<td>version 0.01664</td>
<td>updated 0.02317</td>
<td>finite 0.01242</td>
<td>strategies 0.01123</td>
</tr>
<tr>
<td>bruce 0.01273</td>
<td>enron 0.01067</td>
<td>pdf 0.01421</td>
<td>checked 0.02277</td>
<td>length 0.01154</td>
<td>opponent... 0.01088</td>
</tr>
<tr>
<td>visit 0.01232</td>
<td>dataset 0.00960</td>
<td>bug 0.01332</td>
<td>change 0.02156</td>
<td>strings 0.01013</td>
<td>nash 0.00848</td>
</tr>
</tbody>
</table>
Latent Dirichlet Allocation

Results:
conference proceedings (NIPS)

Relative weights of various topics in 17 years of NIPS proceedings
Further reading:
“Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends”
(Wang & McCallum, 2006)
http://people.cs.umass.edu/~mccallum/papers/tot-kdd06.pdf
Temporal dynamics of social networks
How can we characterize, model, and reason about the structure of social networks?

1. Models of network structure
2. Power-laws and scale-free networks, “rich-get-richer” phenomena
3. Triadic closure and “the strength of weak ties”
4. Small-world phenomena
5. Hubs & Authorities; PageRank
Two weeks ago we saw some processes that model the generation of social and information networks

- Power-laws & small worlds
- Random graph models

These were all defined with a “static” network in mind. But if we observe the order in which edges were created, we can study how these phenomena change as a function of time.

First, let’s look at “microscopic” evolution, i.e., evolution in terms of individual nodes in the network.
Temporal dynamics of social networks

Q1: How do networks grow in terms of the number of nodes over time?

(From Leskovec, 2008 (CMU Thesis))

A: Doesn’t seem to be an obvious trend, so what do networks have in common as they evolve?
**Q2:** When do nodes create links?
- x-axis is the age of the nodes
- y-axis is the number of edges created at that age

**A:** In most networks there's a “burst” of initial edge creation which gradually flattens out. Very different behavior on LinkedIn (guesses as to why?)
Q3: How long do nodes “live”?
• x-axis is the diff. between date of last and first edge creation
  • y-axis is the frequency

A: Node lifetimes follow a power-law: many many nodes are shortlived, with a long-tail of older nodes
Temporal dynamics of social networks

What about “macroscopic” evolution, i.e., how do global properties of networks change over time?

**Q1:** How does the # of nodes relate to the # of edges?

- A few more networks: citations, authorship, and autonomous systems (and some others, not shown)
- **A:** Seems to be linear (on a log-log plot) but the number of edges grows faster than the number of nodes as a function of time
Temporal dynamics of social networks

Q1: How does the # of nodes relate to the # of edges?

A: seems to behave like

\[ E(t) \propto N(t)^a \]

where

\[ 1 \leq a \leq 2 \]

• a = 1 would correspond to constant out-degree – which is what we might traditionally assume
• a = 2 would correspond to the graph being fully connected
• What seems to be the case from the previous examples is that a > 1 – the number of edges grows faster than the number of nodes
Q2: How does the degree change over time?

- **A:** The average out-degree increases over time.
Q3: If the network becomes **denser**, what happens to the (effective) diameter?

**A:** The diameter seems to decrease.

In other words, the network becomes more of a small world as the number of nodes increases.
Q4: Is this something that must happen – i.e., if the number of edges increases faster than the number of nodes, does that mean that the diameter must decrease?

A: Let’s construct random graphs (with \(a > 1\)) to test this:
So, a decreasing diameter is not a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model.

**Q5:** is the degree distribution of the nodes sufficient to explain the observed phenomenon?

**A:** Let’s perform random rewiring to test this.

Random rewiring preserves the degree distribution, and randomly samples amongst networks with observed degree distribution.
So, a decreasing diameter is not a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model.

Q5: is the degree distribution of the nodes sufficient to explain the observed phenomenon?
So, a decreasing diameter is **not** a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model.

**Q5:** is the degree distribution of the nodes sufficient to explain the observed phenomenon?

**A:** Yes! The fact that real-world networks seem to have decreasing diameter over time can be explained as a result of their degree distribution **and** the fact that the number of edges grows faster than the number of nodes.
Temporal dynamics of social networks

Other interesting topics...

“memetracker”
Temporal dynamics of social networks

Other interesting topics...

Aligning query data with disease data – Google flu trends:
https://www.google.org/flutrends/us/#US

Sodium content in recipe searches vs. # of heart failure patients – “From Cookies to Cooks” (West et al. 2013):
Further reading:
“Dynamics of Large Networks” (most plots from here)
Jure Leskovec, 2008

“Microscopic Evolution of Social Networks”
Leskovec et al. 2008

“Graph Evolution: Densification and Shrinking Diameters”
Leskovec et al. 2007
CSE 158 – Lecture 17
Web Mining and Recommender Systems

Some incredible assignments
Fake news detection

Grab real and fake news from Kaggle (fake news detection dataset) and Freedom to Tinker (real headlines):

<table>
<thead>
<tr>
<th>Website</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>abcnews.go.com</td>
<td>55486</td>
</tr>
<tr>
<td>bbc.co.uk</td>
<td>37258</td>
</tr>
<tr>
<td>breitbart.com</td>
<td>148836</td>
</tr>
<tr>
<td>buzzfeed.com</td>
<td>110848</td>
</tr>
<tr>
<td>cbsnews.com</td>
<td>87849</td>
</tr>
<tr>
<td>chicagotribune.com</td>
<td>33884</td>
</tr>
<tr>
<td>chron.com</td>
<td>142965</td>
</tr>
<tr>
<td>cnbc.com</td>
<td>38995</td>
</tr>
<tr>
<td>cnn.com</td>
<td>74237</td>
</tr>
<tr>
<td>forbes.com</td>
<td>28077</td>
</tr>
<tr>
<td>foxnews.com</td>
<td>104173</td>
</tr>
<tr>
<td>hollywoodreporter.com</td>
<td>36237</td>
</tr>
<tr>
<td>huffingtonpost.com</td>
<td>72268</td>
</tr>
<tr>
<td>latimes.com</td>
<td>139728</td>
</tr>
<tr>
<td>money.cnn.com</td>
<td>17591</td>
</tr>
<tr>
<td>nbcnews.com</td>
<td>57621</td>
</tr>
<tr>
<td>nytimes.com</td>
<td>171325</td>
</tr>
<tr>
<td>politico.com</td>
<td>18462</td>
</tr>
<tr>
<td>reuters.com</td>
<td>64474</td>
</tr>
<tr>
<td>theguardian.com</td>
<td>68684</td>
</tr>
<tr>
<td>time.com</td>
<td>199723</td>
</tr>
<tr>
<td>usatoday.com</td>
<td>25613</td>
</tr>
<tr>
<td>usnews.com</td>
<td>11859</td>
</tr>
<tr>
<td>wsj.com</td>
<td>61391</td>
</tr>
</tbody>
</table>

Words from real vs. fake headlines

Extract words and train using a CNN

Jimmy Gia Quach, Shih-Cheng Huang
Anime Recommendation

MyAnimeList dataset from Kaggle

$$r = \frac{\sum_{v \in V} \text{similarity}(u, v) \cdot \text{rating}(v, a)}{\text{count}(a)}$$

<table>
<thead>
<tr>
<th>Features</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always predict average</td>
<td>1.03293792426</td>
</tr>
<tr>
<td>Synopsis bag-of-words</td>
<td>0.806018062926</td>
</tr>
<tr>
<td>Genre, members, title</td>
<td>0.681102399363</td>
</tr>
<tr>
<td>All of the above</td>
<td>0.62533064608</td>
</tr>
</tbody>
</table>
Beer reviews

Yunsheng Li, Mengzhi Li, Chenxi Cao
Used car price prediction

Price vs. registration year

Price vs. mileage
- Type (sedan, van, etc.)
- Mileage
- Age
- PowerPS
- Damage
- Gearbox
- Fuel type

Kaggle used cars dataset (370,000 instances)

Xinyuan Zhang, Changtong Qiu, Zhiye Zhang
Death clock

CDC Mortality Dataset (2.1 million instances)

<table>
<thead>
<tr>
<th>Models</th>
<th>Feature List</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>Linear Regression using: {All above features} and 2,100,000 deaths</td>
<td>142.43</td>
</tr>
<tr>
<td>2nd</td>
<td>Linear Regression using: {All above features except Activity Code features} and 2,100,000 deaths</td>
<td>142.89</td>
</tr>
<tr>
<td>3rd</td>
<td>Linear Regression using: {All above features except Resident Code features} and 2,100,000 deaths</td>
<td>143.52</td>
</tr>
<tr>
<td>4th</td>
<td>Linear Regression using: {All above features except Education Level features} and 2,100,000 deaths</td>
<td>144.24</td>
</tr>
<tr>
<td>5th</td>
<td>Linear Regression using: {All above features except Marital Status features} and 2,100,000 deaths</td>
<td>186.49</td>
</tr>
<tr>
<td>6th</td>
<td>SVM using: {All above features} and 50,000 deaths</td>
<td>178.20</td>
</tr>
<tr>
<td>Baseline</td>
<td>Mean age at death</td>
<td>270.25</td>
</tr>
</tbody>
</table>

Daphne Angeline Gunawan, Brandon Jihwan Hwang, Alan Yian Xu, Franklin Alexander Velasquez
Uber pickups

NYC Uber Dataset (14.2 million samples)

Lilith Huang, Aamir Abdur Rasheed
Rental recommendations

#bathrooms

distance to city center

Interest level:

Wen Zhang, Xingbo Wang, Kaixiang Zhao, Lifan Chen Shiunn An Lu, Shanyu Chuang, Hao-En Sung Side Li, Yifan Xu Dhruv Sharma, Keshav Sharma, Saransh Jain
Crime prediction

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Unique identifier for the record</td>
</tr>
<tr>
<td>Case Number</td>
<td>The Chicago Police Department RD Number</td>
</tr>
<tr>
<td>Date</td>
<td>Date when the incident occurred in mm/dd/yyyy.</td>
</tr>
<tr>
<td>Block</td>
<td>The partially redacted address where the incident occurred.</td>
</tr>
<tr>
<td>IUCR</td>
<td>The Illinois Uniform Crime Reporting code</td>
</tr>
<tr>
<td>Primary Type</td>
<td>The primary description of the IUCR code</td>
</tr>
<tr>
<td>Description</td>
<td>The secondary description of the IUCR code</td>
</tr>
<tr>
<td>Location</td>
<td>Description of the location where the incident occurred.</td>
</tr>
<tr>
<td>Arrest</td>
<td>Indicated whether an arrest was made.</td>
</tr>
<tr>
<td>Domestic</td>
<td>Indicates whether the incident was domestic-related</td>
</tr>
<tr>
<td>Beat</td>
<td>Indicates the beat (the smallest police geographic area) where the incident occurred.</td>
</tr>
<tr>
<td>District</td>
<td>Indicates the police district where the incident occurred.</td>
</tr>
<tr>
<td>Ward</td>
<td>The ward (City Council district) where the incident occurred.</td>
</tr>
<tr>
<td>Community Area</td>
<td>Indicates the community area where the incident occurred. Chicago has 77 community areas.</td>
</tr>
<tr>
<td>FBI Code</td>
<td>Indicates the crime classification as outlined in the FBI’s National Incident-Based Reporting System (NIBRS).</td>
</tr>
<tr>
<td>X Coordinate</td>
<td>The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.</td>
</tr>
<tr>
<td>Y Coordinate</td>
<td>The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.</td>
</tr>
<tr>
<td>Year</td>
<td>Year the incident occurred.</td>
</tr>
<tr>
<td>Updated On</td>
<td>Date and time the record was last updated.</td>
</tr>
<tr>
<td>Latitude</td>
<td>The latitude of the location where the incident occurred.</td>
</tr>
<tr>
<td>Longitude</td>
<td>The longitude of the location where the incident occurred.</td>
</tr>
</tbody>
</table>

Crime types by hour

- Theft by location
  - Theft
    - Day
    - Year

- Narcotics

Wenbin Zhu, Yuchen Wang, Wenjie Tao
Sahil Agarwal, Ujjwal Gulecha, Shalini Kedlaya
Junyang Li, Shenghong Wang
H1B petitions

Kaggle dataset (~1 million samples)

Fig. 8. State-wise Median wage

<table>
<thead>
<tr>
<th>Job title</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROGRAMMER ANALYST</td>
<td>TATA CONSULTANCY</td>
</tr>
<tr>
<td>SOFTWARE ENGINEER</td>
<td>INFOSYS</td>
</tr>
</tbody>
</table>

Yuchen Feng, Xuanzhen Xu, Jianxiong Lin
Prahal Arora, Rahul Vijay Dubey, Induja Sreekanthan, Jahnvi Singhal
Jialin Wang, Yishu Ma, Han Li
Kobe field goals

Kaggle competition of 30,000 field-goal attempts
Taxi tips

Rushil Nagda, Sudhanshu Bahety, Shubham Gupta, Tejas Saxena, Himanshu Jaiswal, Tushar Bansal, Prateek Ravindra Jakate
Fill out those evaluations!

- Please evaluate the course on http://cape.ucsd.edu/students!
Thanks!