NLP and Information Retrieval
자연어처리 기초

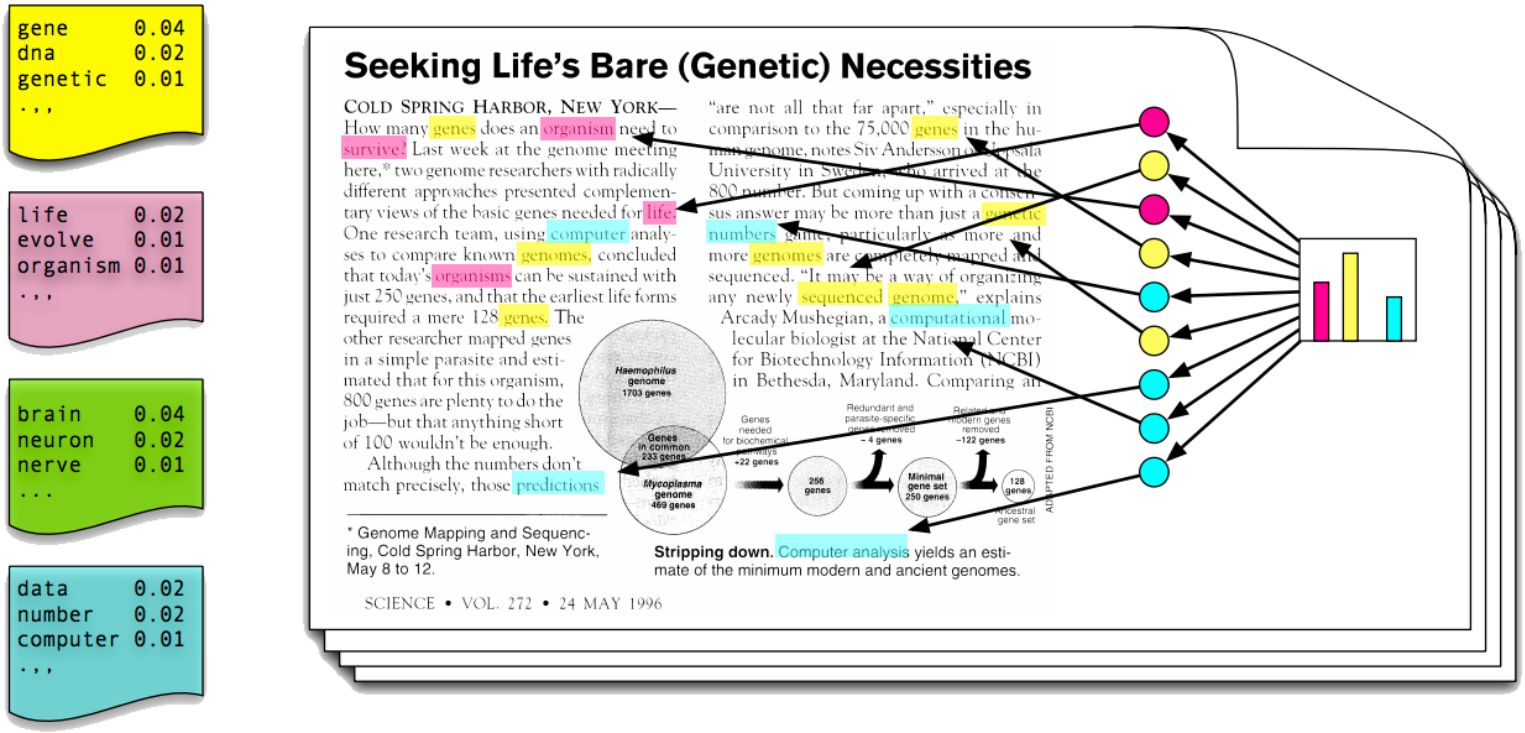
Prof. Joongheon Kim
School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea
https://sites.google.com/site/joongheonkim/
joongheon@gmail.com

자연어처리 기본가정

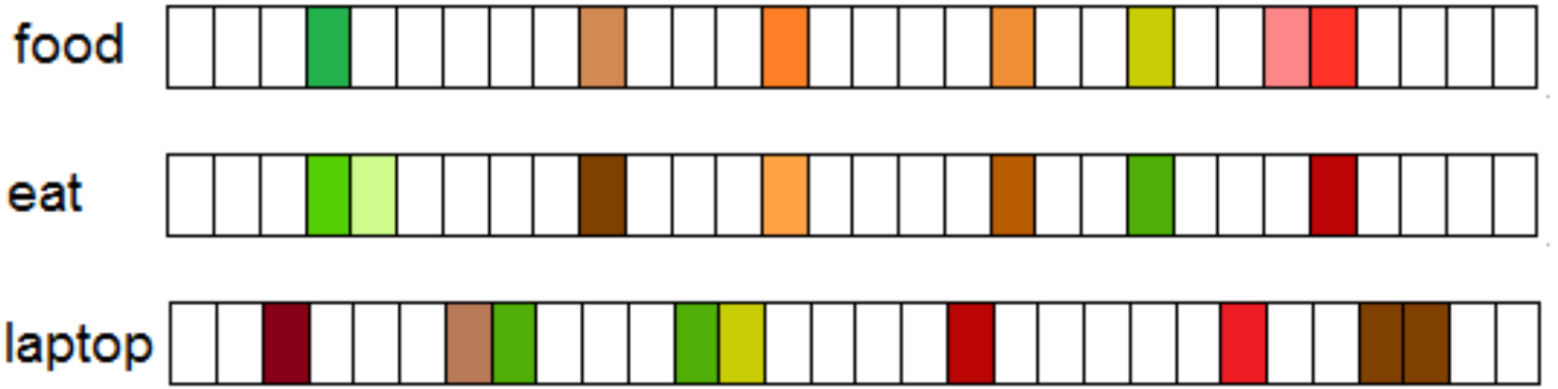
- NLP의 기본 가정
  - **Distributional Hypothesis:** 비슷한 맥락에 등장하는 단어들은 유사한 의미를 지니는 경향이 있다. (words that occur in similar contexts tend to have similar meanings)
  - **Vector Space Models:** 문서 집합에 속하는 각각의 문서들을 벡터공간의 벡터로 표현(representation)할 수 있다. 벡터공간에 벡터로 표현된 문서들 사이의 거리가 가깝다면 의미가 유사하다 (semantically similar).
  - **Statistical Semantics Hypothesis:** 언어 사용의 통계적 패턴은 사람들이 의미하는 바를 이해하는 데 쓰일 수 있다. (statistical patterns of human word usage can be used to figure out what people mean)
  - **Bag of Words Hypothesis:** 어떤 문서에 출현한 단어들의 빈도는 문서와 쿼리의 관련성을 나타내는 경향이 있다. (the frequencies of words in a document tend to indicate the relevance of the document to a query) 어떤 문서가 쿼리 문서와 유사한 벡터라면 그 의미도 비슷하다.
  - **Latent Relation Hypothesis:** 비슷한 패턴으로 동시에 등장하는 단어쌍은 유사한 의미적 관계를 지니는 경향이 있다. (Pairs of words that co-occur in similar patterns tend to have similar semantic relations)

- Topic Modeling (Latent Dirichlet Allocation)

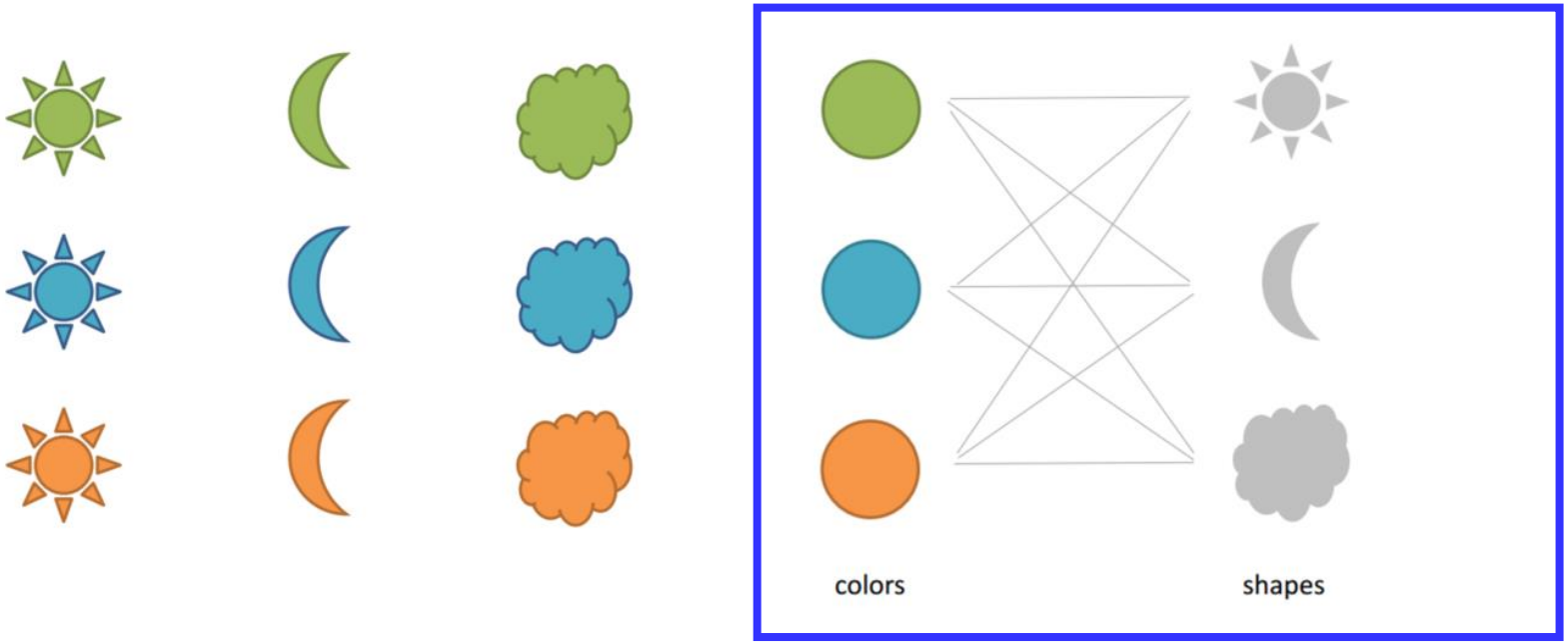
Topics *Documents* *Topic proportions & assignments*



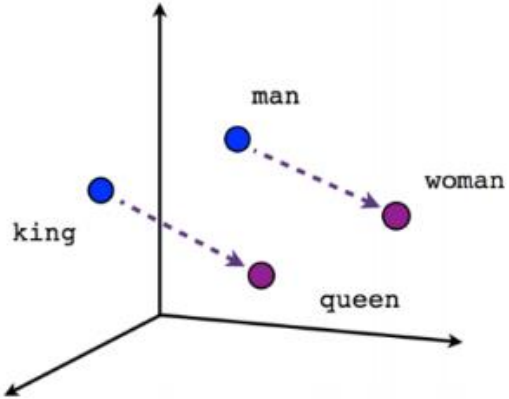
- Word Embedding
  - Based on distributional hypothesis



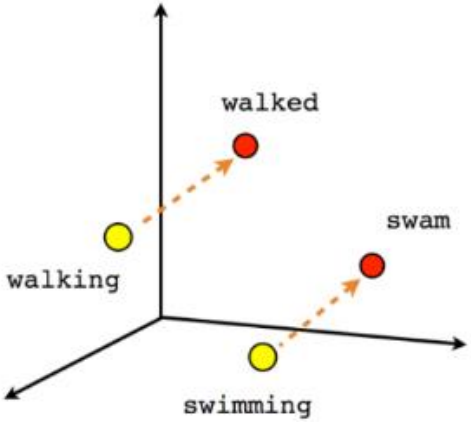
- Word Embedding
  - Based on distributional hypothesis



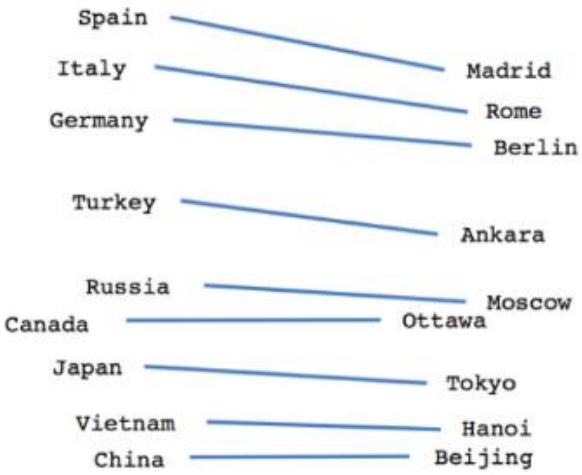
- Word Embedding
  - Based on distributional hypothesis



Male-Female

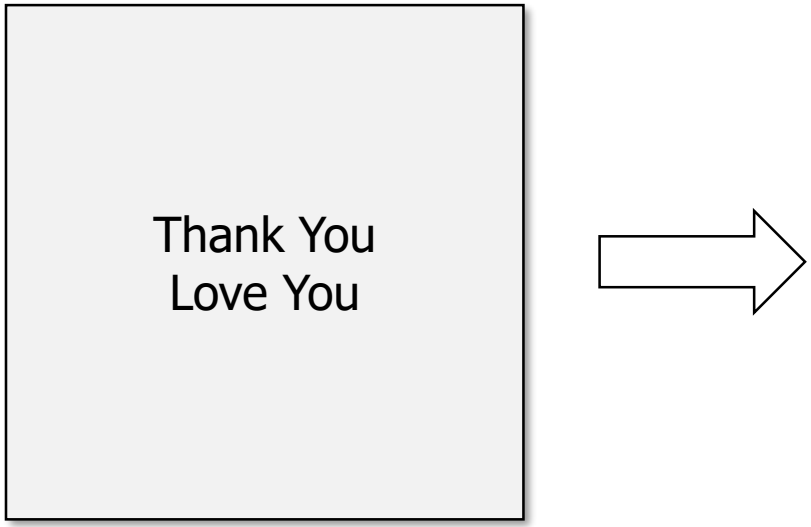


Verb tense



Country-Capital

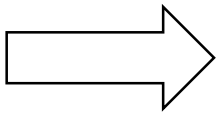
- What is encoding?
  - Convert text to **number**



Unique word	Encoding
Thank	0
You	1
Love	2

- What is one hot encoding?
  - Convert text to **vector**

	Thank	You	Love
Thank	1	0	0
You	0	1	0
Love	0	0	1



Unique word	Encoding
Thank	[1,0,0]
You	[0,1,0]
Love	[0,0,1]

- One hot encoding
  - However,
    - One hot encoding does not have **similarity**
    - Every distance is same to each other
    - Cosine similarity also 0 since angle is 90 degree

Unique word	Encoding
Thank	[1,0,0]
You	[0,1,0]
Love	[0,0,1]

- Embedding
  - Embedding is dense vector with similarity

Unique word	Encoding	Embedding
King	[1,0,0,0]	[1,2]
Man	[0,1,0,0]	[1,3]
Queen	[0,0,1,0]	[5,1]
Woman	[0,0,0,1]	[5,2]

자연어처리 기초: Word2Vec

- Definition of Word2Vec
  - Word2Vec is word embedding
  - Similarity comes from neighbor words

- Word2Vec data generation (skipgram)
  - Windows size: 1
    - King Brave Man
    - Queen Beautiful Woman

Word	Neighbor
King	Brave
Brave	King
Brave	Man
Man	Brave
Queen	Beautiful
Beautiful	Queen
Beautiful	Woman
Woman	Beautiful

- Word2Vec data generation (skipgram)
  - Windows size: 2
    - King Brave Man
    - Queen Beautiful Woman

Word	Neighbor
King	Brave
King	Man
Brave	King
Brave	Man
Man	King
Man	Brave
Queen	Beautiful
Queen	Woman
Beautiful	Queen
Beautiful	Woman
Woman	Queen
Woman	Beautiful

자연어처리 기초: Word2Vec

- Word2Vec data generation (skipgram), window size: 2

Word	Word (One-Hot)	Neighbor	Neighbor (One-Hot)
King	[1,0,0,0,0,0]	Brave	[0,1,0,0,0,0]
King	[1,0,0,0,0,0]	Man	[0,0,1,0,0,0]
Brave	[0,1,0,0,0,0]	King	[1,0,0,0,0,0]
Brave	[0,1,0,0,0,0]	Man	[0,0,1,0,0,0]
Man	[0,0,1,0,0,0]	King	[1,0,0,0,0,0]
Man	[0,0,1,0,0,0]	Brave	[0,1,0,0,0,0]
Queen	[0,0,0,1,0,0]	Beautiful	[0,0,0,0,1,0]
Queen	[0,0,0,1,0,0]	Woman	[0,0,0,0,0,1]
Beautiful	[0,0,0,0,1,0]	Queen	[0,0,0,1,0,0]
Beautiful	[0,0,0,0,1,0]	Woman	[0,0,0,0,0,1]
Woman	[0,0,0,0,0,1]	Queen	[0,0,0,1,0,0]
Woman	[0,0,0,0,0,1]	Beautiful	[0,0,0,0,1,0]

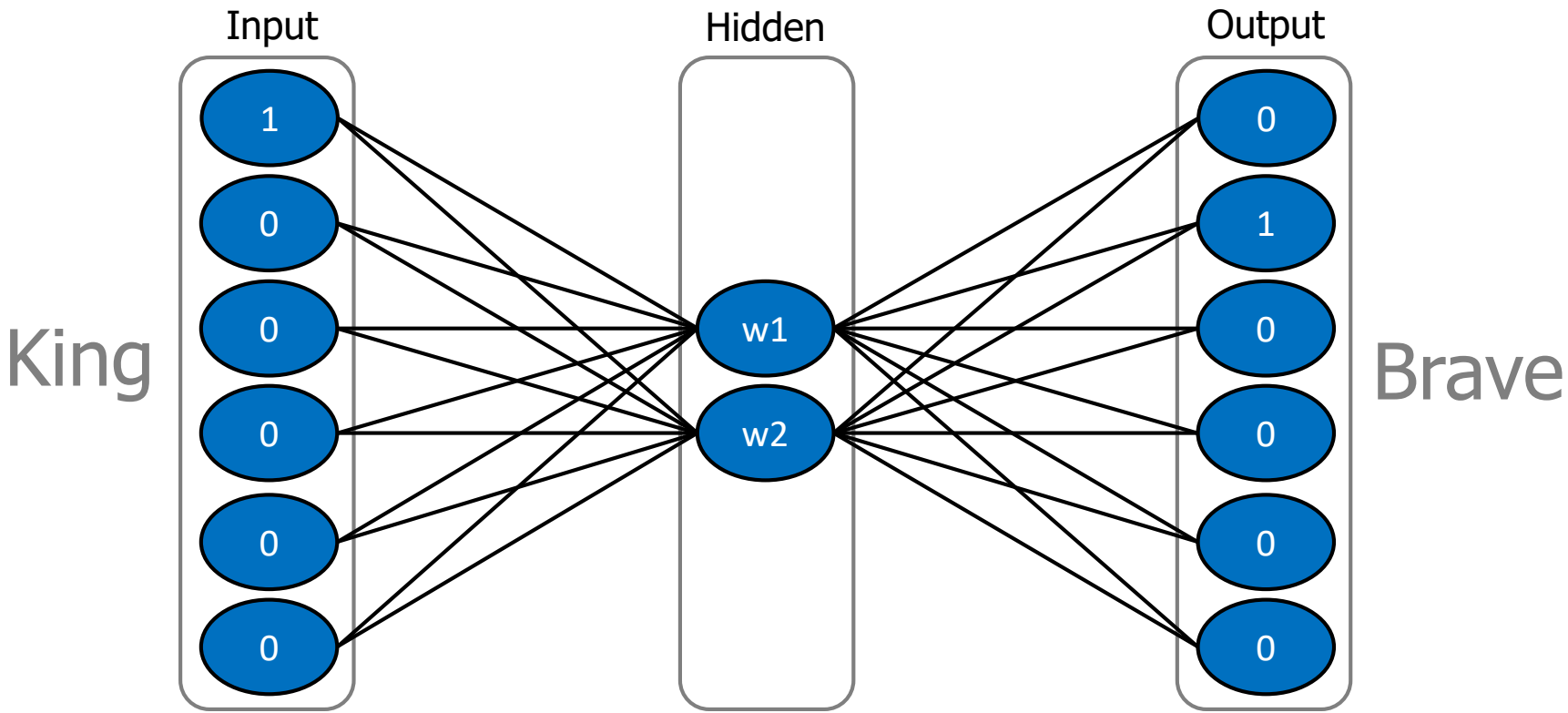
- Word2Vec data generation (skipgram), window size: 2

Word	Word (One-Hot)	Neighbor	Neighbor (One-Hot)
King	[1,0,0,0,0,0]	Brave	[0,1,0,0,0,0]
King	[1,0,0,0,0,0]	Man	[0,0,1,0,0,0]
Brave	[0,1,0,0,0,0]	King	[1,0,0,0,0,0]
Brave	[0,1,0,0,0,0]	Man	[0,0,1,0,0,0]
Man	[0,0,1,0,0,0]	King	[1,0,0,0,0,0]
Man	[0,0,1,0,0,0]	Brave	[0,1,0,0,0,0]
Queen	[0,0,0,1,0,0]	Beautiful	[0,0,0,0,1,0]
Queen	[0,0,0,1,0,0]	Woman	[0,0,0,0,0,1]
Beautiful	[0,0,0,0,1,0]	Queen	[0,0,0,1,0,0]
Beautiful	[0,0,0,0,1,0]	Woman	[0,0,0,0,0,1]
Woman	[0,0,0,0,0,1]	Queen	[0,0,0,1,0,0]
Woman	[0,0,0,0,0,1]	Beautiful	[0,0,0,0,1,0]

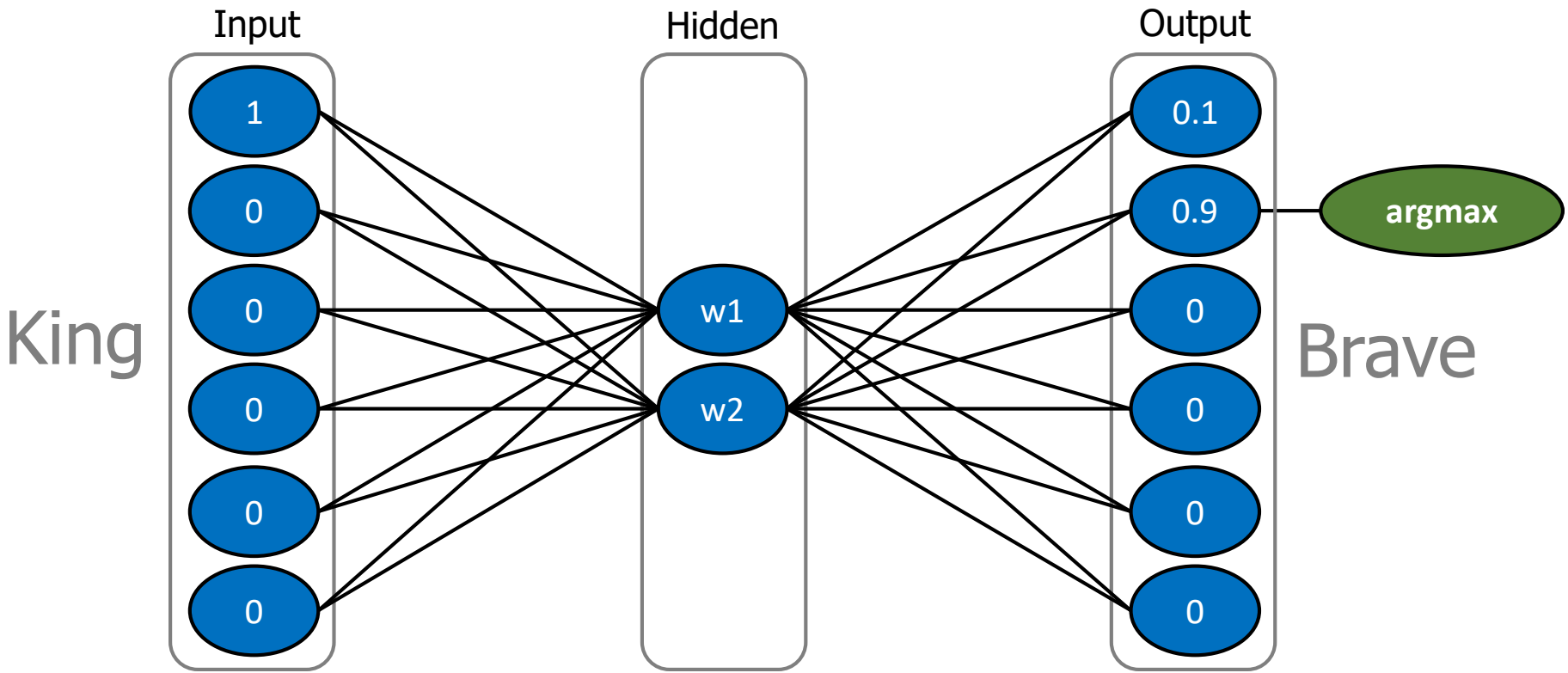
INPUT

OUTPUT

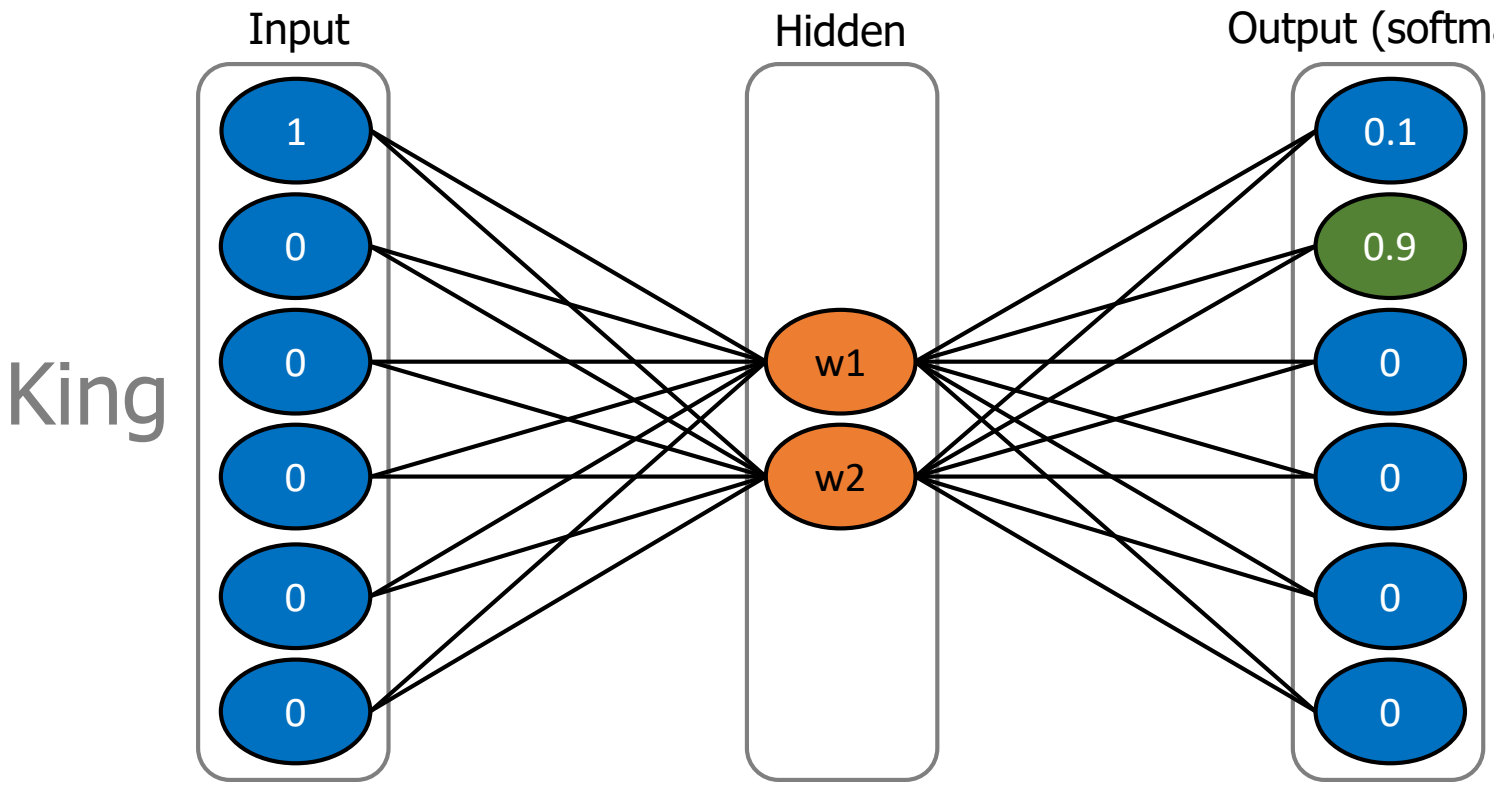
- Word2Vec **training** (the first entry example → input: King, output: Brave)



- Word2Vec **testing**



- Word2Vec

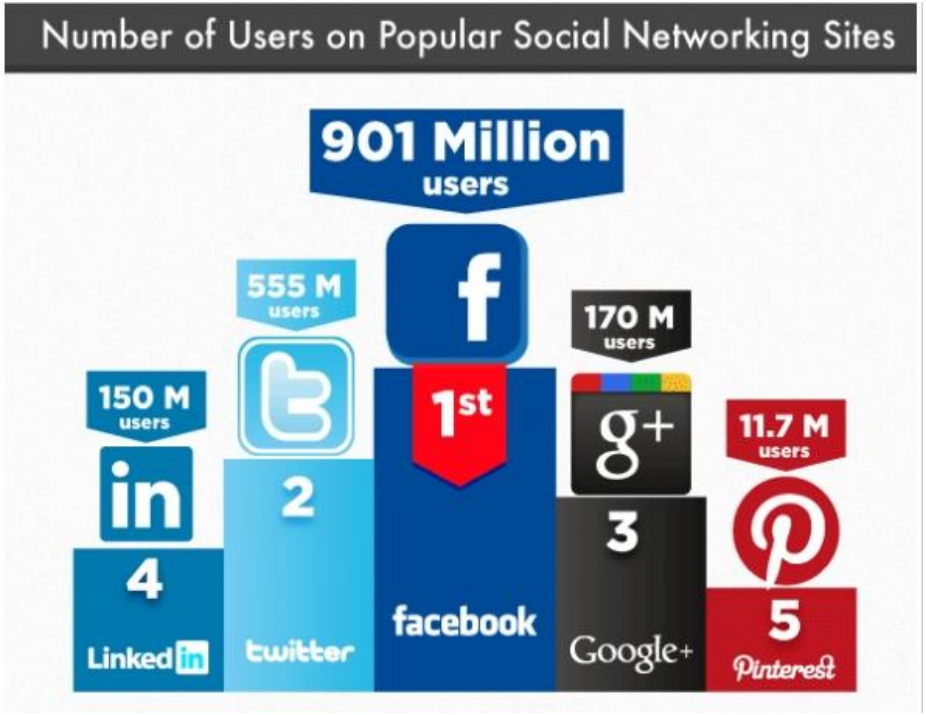


Unique Word	Embedding (w1,w2)
King	[1,1]
Brave	[1,2]
Man	[1,3]
Queen	[5,4]
Beautiful	[5,5]
Woman	[5,6]

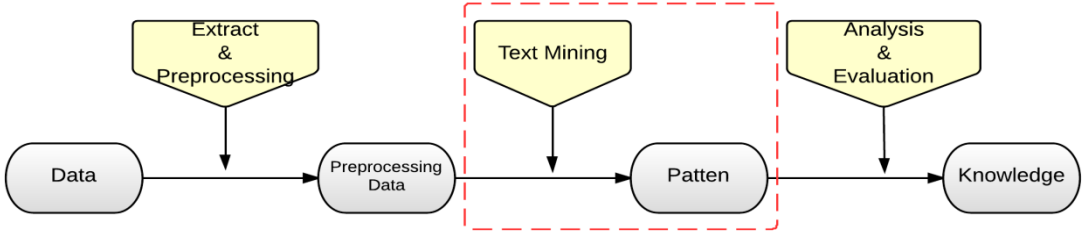
NLP and Information Retrieval
텍스트분석 기초

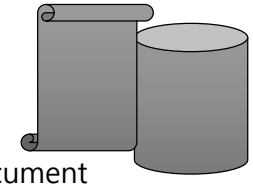
Prof. Joongheon Kim
School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea
https://sites.google.com/site/joongheonkim/
joongheon@gmail.com


25억
2013년
20억
2010년
10억
2005년
4억
2000년
0
1995년
인터넷 사용자 수 추이
HTML
5
WWW


Number of Users on Popular Social Networking Sites
901 Million users
1st
facebook
555 M users
2
twitter
150 M users
4
LinkedIn
170 M users
3
Google+
11.7 M users
5
Pinterest

- Text Mining Procedure





Document collection

1

Establish the Corpus : Collect and organize the domain-specific data

Corpus

2

Pre-process the Data: Introduce structure to the corpus

Structured data

Patterns & Associations

3

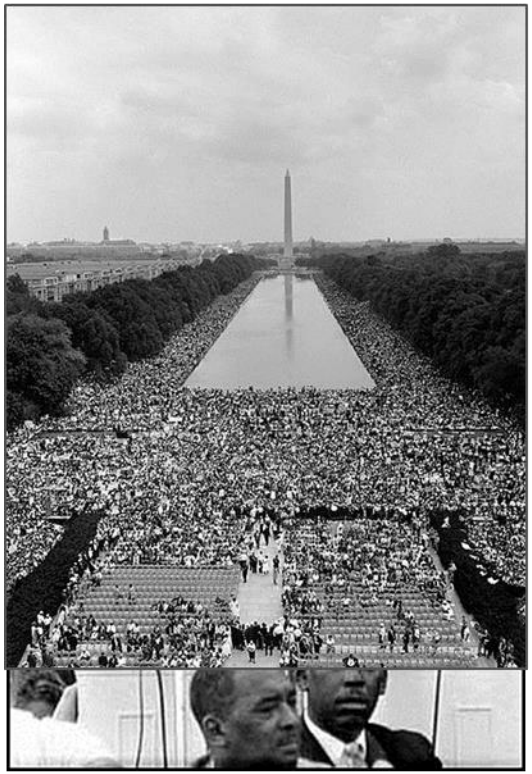
Extract knowledge: Discover patterns from the preprocessed data

Feedback

Feedback

Feedback : adjustment / corrections

- "I Have a Dream" by Martin Luther King



- **Document's Overview**

Sentences	85
Words	1579
Different Words	523
Words per sentence	18.57

King's speech invokes
- ✓ the Declaration of Independence
- ✓ the Emancipation Proclamation
- ✓ the United States Constitution

Ranked **the top American speech** of the 20th century by a 1999 poll of scholars of public address

텍스트분석 예제

- Keywords

I HAVE A DREAM

Martin Luther King Jr.

I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation.

Five score years ago, a great American, in whose symbolic shadow we stand today, signed the Emancipation Proclamation. This momentous decree came as a great beacon light of hope to millions of Negro slaves who had been seared in the flames of withering injustice. It came as a joyous daybreak to end the long night of their captivity.

But one hundred years later, the Negro still is not free. One hundred years later, the life of the Negro is still sadly crippled by the manacles of segregation and the chains of discrimination. One hundred years later, the Negro lives on a lonely island of poverty in the midst of a vast ocean of material prosperity. One hundred years later, the Negro is still languished in the corners of American society and finds himself an exile in his own land. And so we've come here today to dramatize a shameful condition.

In a sense we've come to our nation's capital to cash a check. When the architects of our republic wrote the magnificent words of the Constitution and the Declaration of Independence, they were signing a promissory note to which every American was to fall heir. This note was a promise that all men, yes, black men as well as white men, would be guaranteed the "unalienable Rights" of "Life, Liberty and the pursuit of Happiness." It is obvious today that America has defaulted on this promissory note, insofar as her citizens of color are concerned. Instead of honoring this sacred obligation, America has given the Negro people a bad check, a check which has come back marked "insufficient funds."

But we refuse to believe that the bank of justice is bankrupt. We refuse to believe that there are insufficient funds in the great vaults...

America, brother
dream, faith
free, god, hope
justice, land, men
mountain, nation
negro, people
swelter, together

- 16 keywords are selected from 523 keywords (3.1%)
- Utilized TF-IDF method

(Term Frequency – Inverse Document Frequency)

$$TF - IDF = TF(t, d) * IDF(t, D)$$

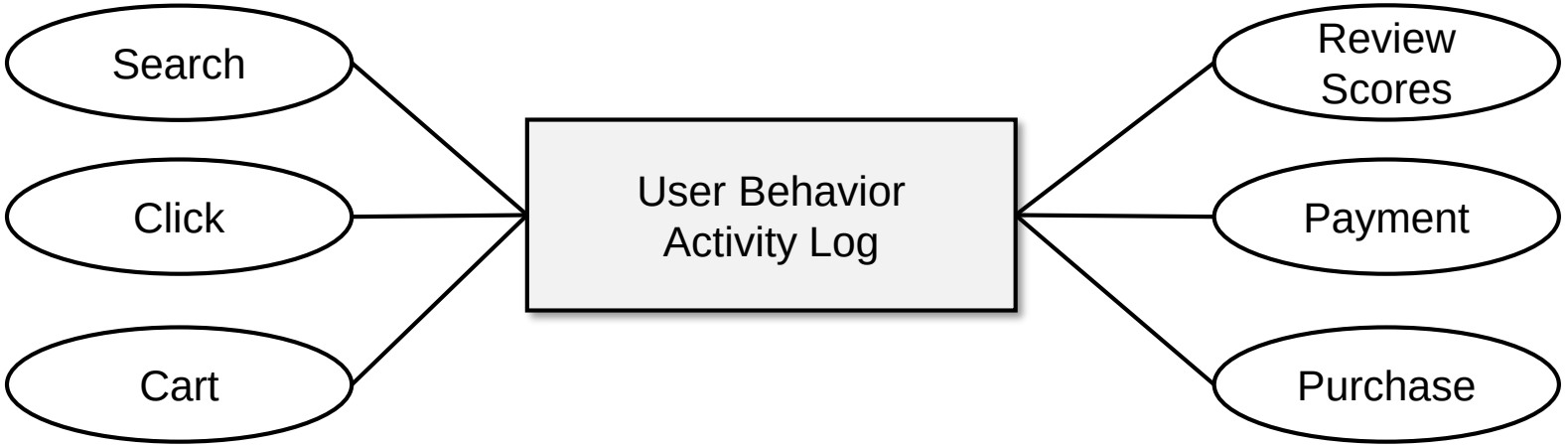
Salton G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill, ISBN 0-07-054484-0.

NLP and Information Retrieval
추천시스템 기초

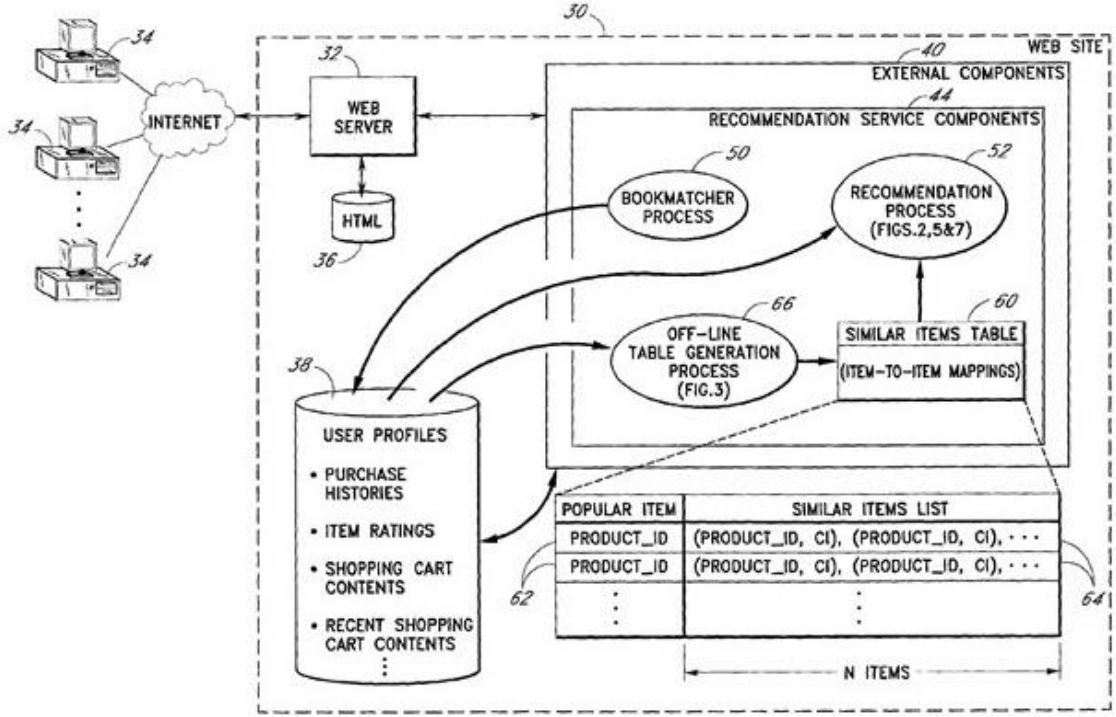
Prof. Joongheon Kim
School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea
https://sites.google.com/site/joongheonkim/
joongheon@gmail.com

추천시스템 기초: 정의

- Recommendation System
  - Based on the user's behavior activity, relations, item similarity, and user contexts, realizing the automated estimation of user preferred items.



- Amazon Recommendation System A9
  - Personalized Recommendation via Item Similarity DB



- Collaborative Filtering (CF)
  - Based on the behavior activity log, conducting the mining of purchase patterns
  - Measuring the similarity for (i) user-to-user, (ii) item-to-item, and (iii) user-to-item
    - Similarity measure via **cosine similarity**

$$\text{simil}(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{||\vec{x}|| \times ||\vec{y}||} = \frac{\sum_{i\in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i\in I_x} r_{x,i}^2}\sqrt{\sum_{i\in I_y} r_{y,i}^2}}$$

- (Step 1) Check the purchased item
- (Step 2) Find the similar items via the cosine similarity
- (Step 3) Make a recommendation for the similar items

추천시스템 기초: Collaborative Filtering

- Collaborative Filtering (CF)
  - Pros
    - Implementable with small number of information
    - Reasonable accuracy are reported in practical applications
  - Cons
    - Use of low-density high-dimensional vectors
    - Scalability issue: handling newly added users/items is not easy with CF

추천시스템 기초: Content-based Filtering

- Content-based Filtering (CF)
  - Recommendation based on item-attributes
  - Analyzing the item itself whereas collaborative filtering focuses on user behavior activity logs
  - Since item-attribute is the core of this algorithm, conducting item analysis and item-similarity measure is important. → **TF-IDF method** is used.

- Collaborative Filtering vs. Content-based Filtering (CF)

	Collaborative Filtering	Content-based Filtering
장점	대부분의 경우 추천성능이 좋음 잠재적인 특징을 고려, 보다 다양한 범위의 추천 가능	사용자의 기호를 직접 반영 새로 추가된 아이템도 추천 가능
단점	아직 평가되지 않은 항목은 추천 대상으로 발견되기 어려움 초기 사용자에 대해선 믿을만한 추천이 어려움	명시적으로 표현된 특징만을 다룰 수 있고 질적 부분을 포착하기 어려움 사용자의 선호도/취향을 특정 단어로 표현하기 어려움 추천하는 항목이 비슷한 장르에 머무르는 한계 존재

- Collaborative Filtering vs. Content-based Filtering (CF)

Collaborative Filtering
- 다수 사용자의 평가 정보 활용
- 다양한 범위의 추천 가능
- 사용자 행동로그 등 빅데이터 활용

Content-based Filtering
- 아이템 설명 정보 활용
- 적은 데이터로 추천 가능
- 좁은 범위 추천

새로 추가된 아이템 추천 가능

다양한 범위 추천 가능

유사성, 잠재요소 등을 고려하여
두 알고리즘과 딥러닝의 특징을 결합한 추천 알고리즘 개발