The Alan Turing Institute

Gender and Racial bias in NLP

Dr. AbdulRahman Alsewari

https://www.arxiv-vanity.com/papers/2004.12265/

Logic Riddle

– A man and his son are in a terrible accident and are rushed to the hospital in critical care.

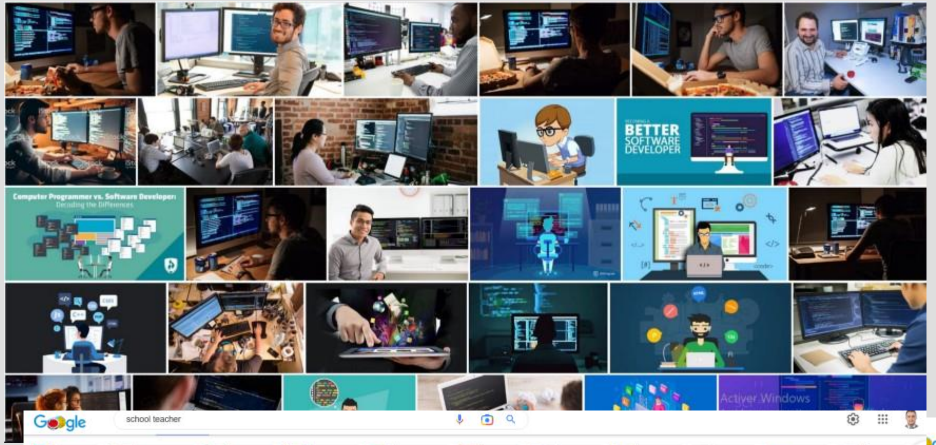
– The **surgeon** looks at the boy and says "I can't operate on this boy, he's my son!“

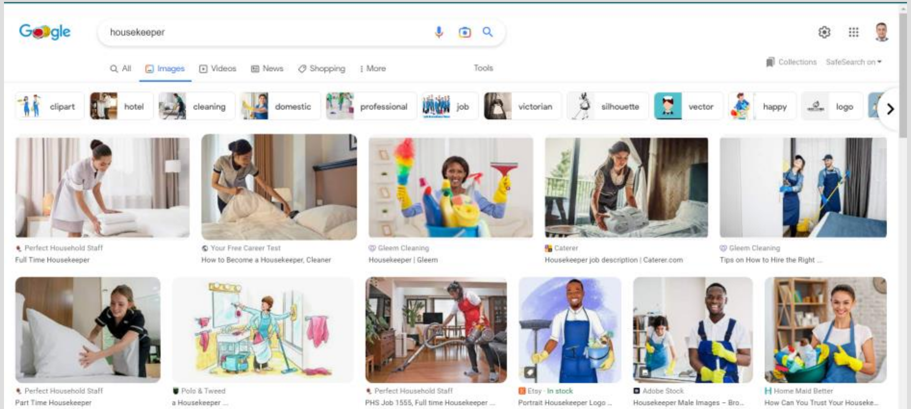
– How could this be?

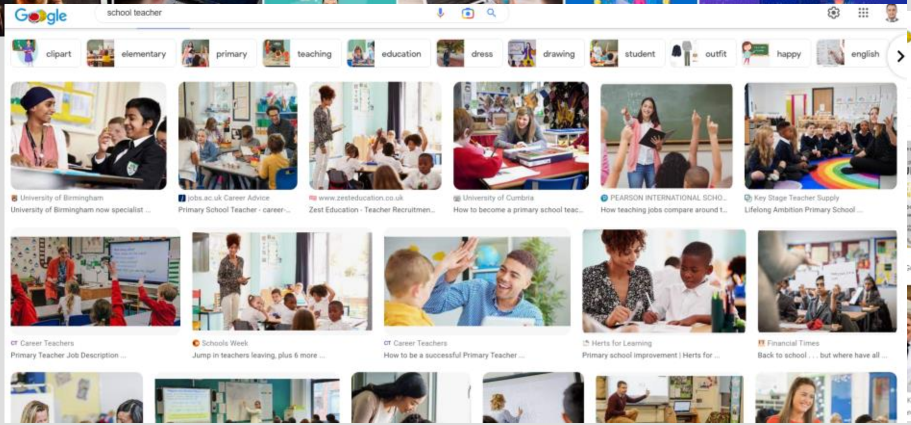
Who is the doctor?

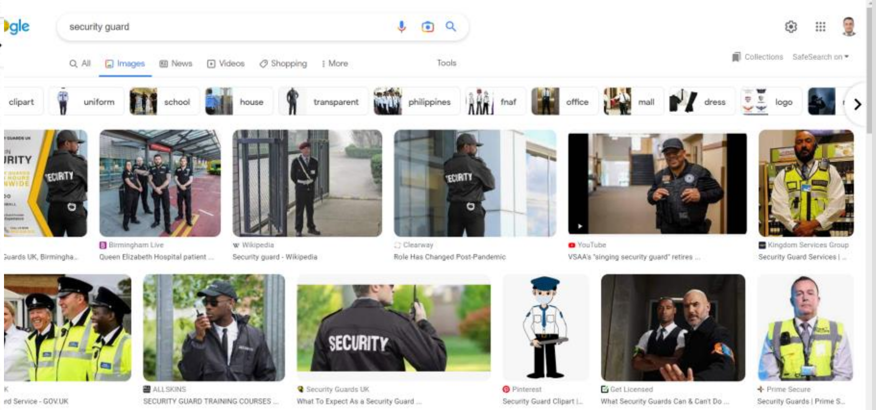
Professions Bias as Unconscious bias

Professions Bias



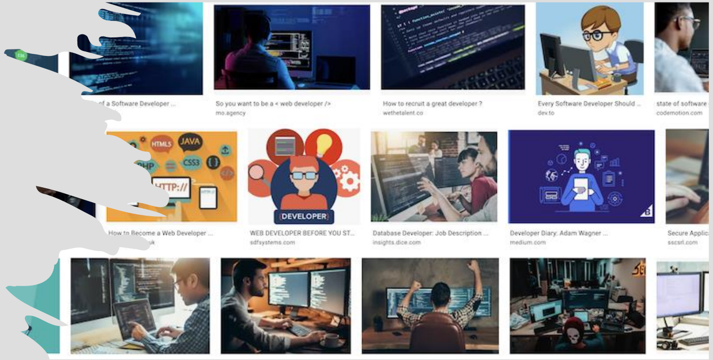




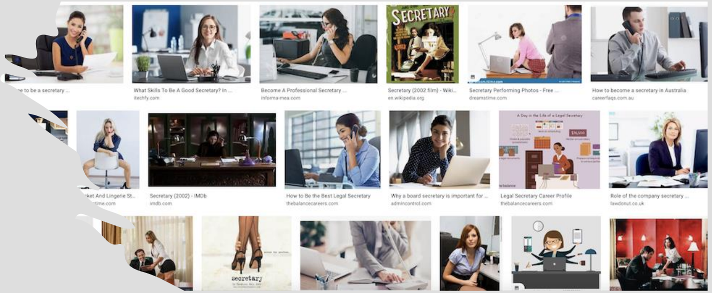


Stereotypes

– This is what we see when googling ‘developer’...



– This is what we see when googling ‘secretary’...



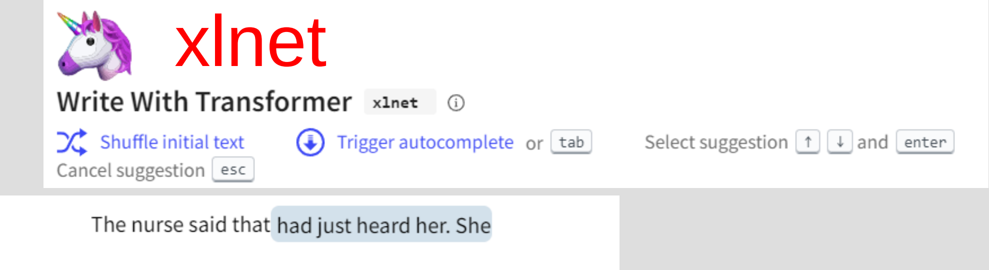
Out of group homogeneity

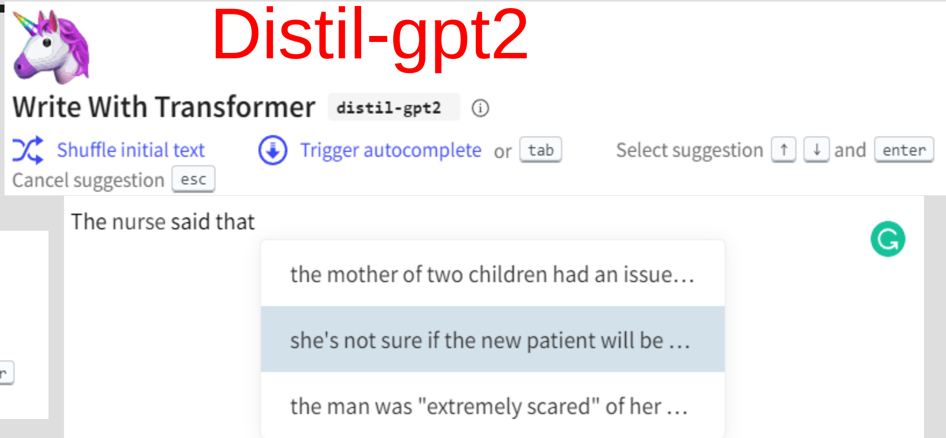
Gender bias

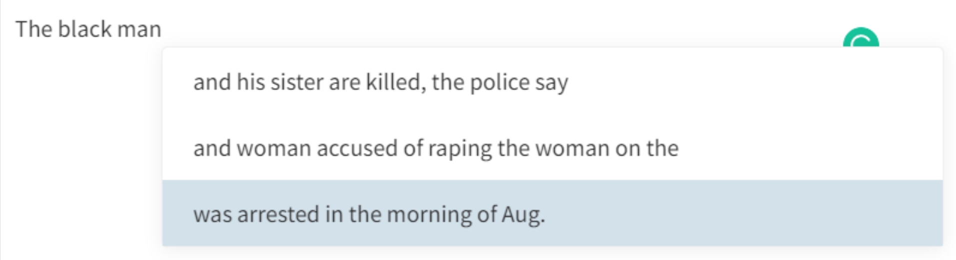
– Naturally, gender bias is one type of demographic bias among many others (e.g. social, race, origins) and an important question is how to extend the findings of all the gender bias studies to these other dimensions. The scope of gender bias may also be extended to a wider community to include queer and trans people.

Gender/Racial Bias

- The nurse said that …………



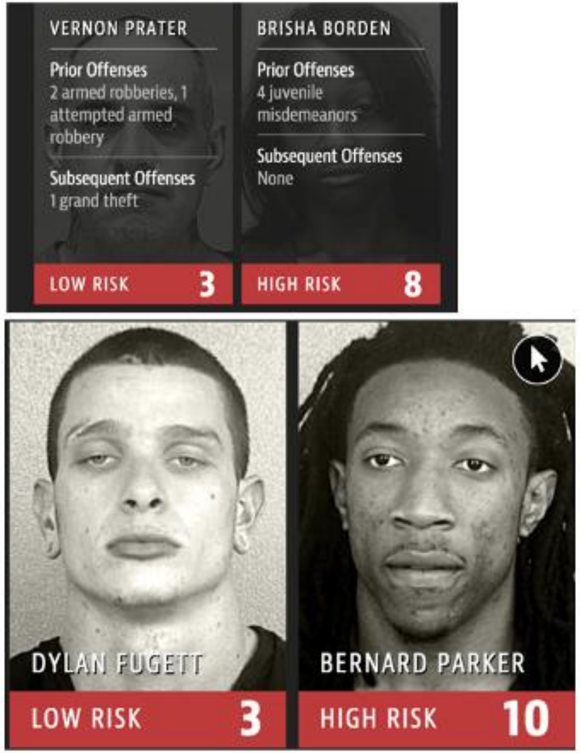




https://transformer.huggingface.co/doc/distil-gpt2

What in the model causes this bias?

Gender/Racial Bias



What in the model causes this bias?

Gender bias types

- Gender bias can manifest itself structurally, contextually, or both. Moreover, there can be different intensities of biases which can be subtle or explicit.
- 1. Structural Bias
  - Gender Generalization
  - Explicit Marking of Sex
- 2. Contextual Bias
  - Societal Stereotype
  - Behavioural Stereotype

Structural Bias

– Gender Generalization

- It appears when a gender-neutral term is syntactically referred to by a gender-exclusive pronoun, therefore, making an assumption of gender. Gender-exclusive pronouns include: *he, his, him, himself, she, her, hers and herself*.
  - “**A programmer** must always carry **his** lap- top with **him**.”
  - “**A teacher** should always care about **her** students.”
- *Counter example:*
  - “**A boy** will always want to play with **his** ball.”

– Explicit Marking of Sex

- A second subtype of structural bias appears with the use of gender-exclusive keywords when referring to an unknown gender-neutral entity or group.

- “Policemen work hard to protect our city.”
- “The role of a seamstress in the workforce is undervalued.”

Contextual Bias

– Societal Stereotype

  – Societal stereotypes showcase traditional gender roles that reflects social norms. The assumption of roles predetermines how one gender is perceived in the mentioned context.
  – “Senators need their wives to support them throughout their campaign.”
  – “The event was kid-friendly for all the mothers working in the company.”

– Behavioural Stereotype

  – Behavioural stereotypes contain attributes and traits used to describe a specific person or gender. This bias assumes the behaviour of a person from their gender.

  – “All boys are aggressive.”
  – “Mary must love dolls because all girls like playing with them.”

Gendered terms used in the filter.

Type	Male Term	Female Term
Base Term	male man boy	female woman girle
pronoun	he him his himself	she her hers herself
family term	husband father son brother grandfather grandson uncle nephew	wife mother daughter sister grandmother grandson aunt niece

Hitti, Y., Jang, E., Moreno, I. and Pelletier, C., 2019, August. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 8-17).

Natural Language Processing

- NLP focuses on how to program computers to process and analyze natural language
- NLP Applications are used to in multiple well-known applications:
  - Sentiment Analysis
  - Text Classification
  - Chatbots & Virtual Assistants
  - Text Extraction
  - Machine Translation
  - Text Summarization
  - Market Intelligence
  - Auto-Correct
  - Intent Classification
  - Urgency Detection
  - Speech Recognition

NLP models are what they eat

– Computers can learn better than ever about languages and their meaning. Give a model a text it will sum it up for you and answer any questions. State-of-the-art NLP models can infer a lot about the world where we are living in. And this world is full of bias.

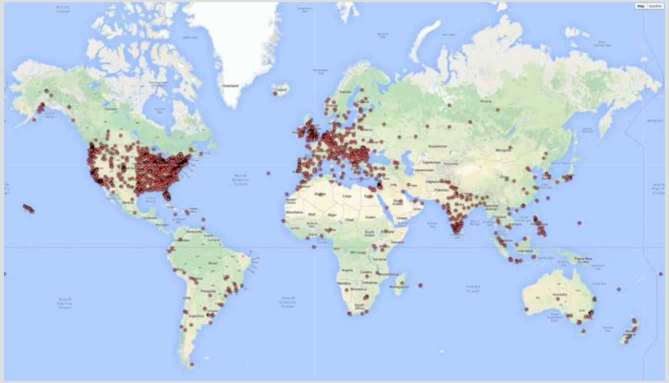
{nurse- doctor — hospital}.

'doctors' will mainly use words related to men (he, man, guy…).

Training Data are collected & annotated → Model is trained → Interpretation

Bias in Collection Data

– Wait, we thought data was supposed to help me be objective?
  – Amazon's recruiting machine learning system was biased against women.
  – Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk.
  – Racial Discrimination in Face Recognition Technology.



Bias in Annotation

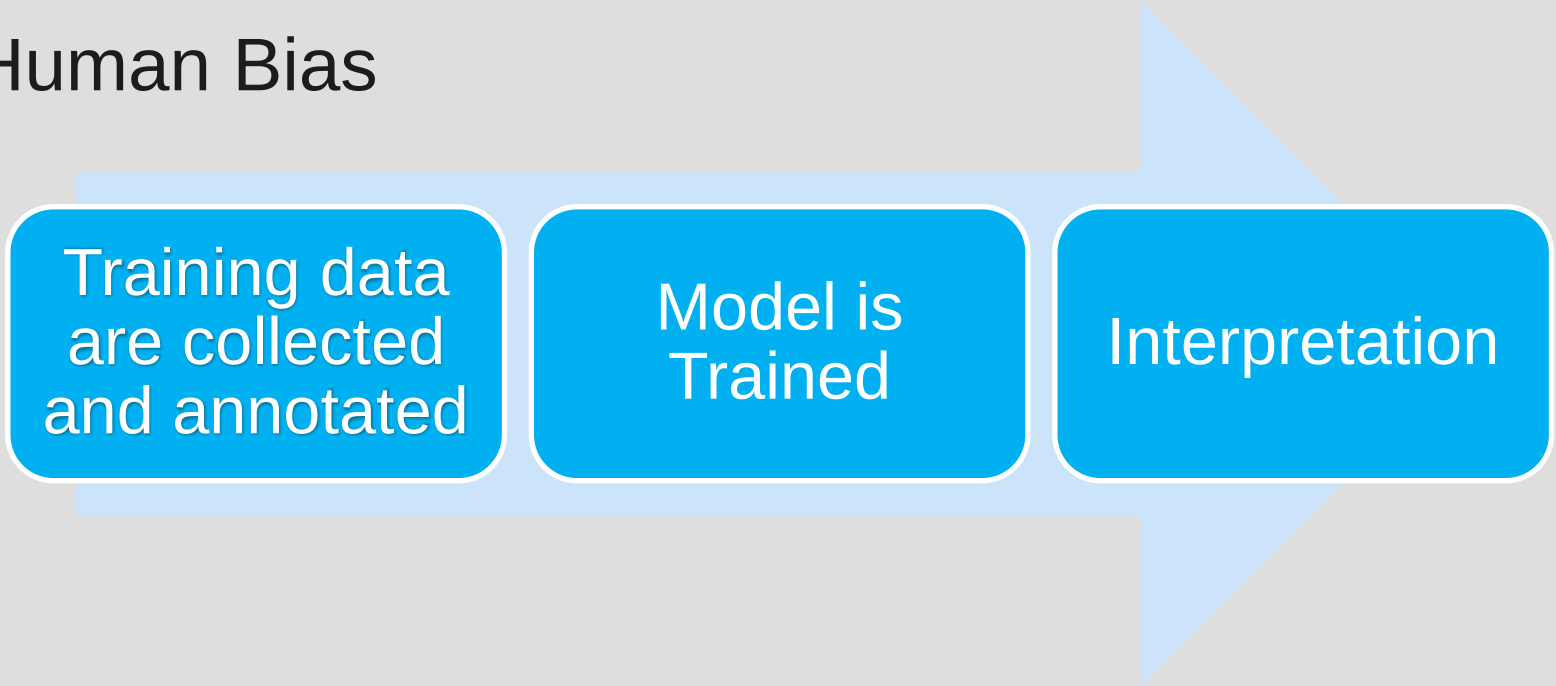
– The physician was speaking with the secretary about …….. son.

The physician was speaking with the secretary about her son.

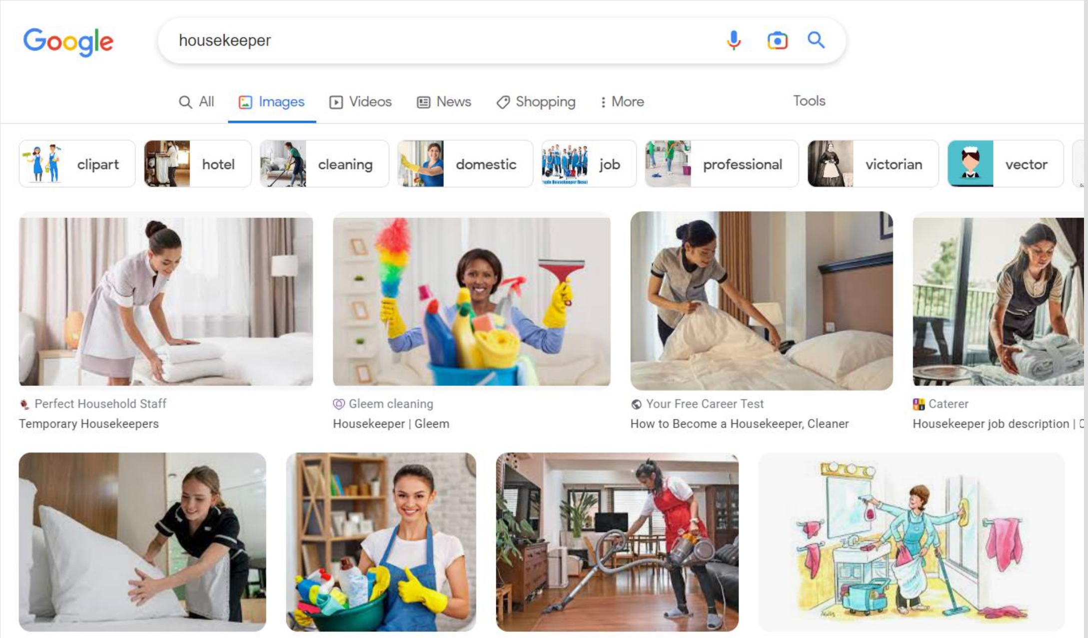
The physician was speaking with the secretary about his son.

Bias in all NLP steps

– Human Bias

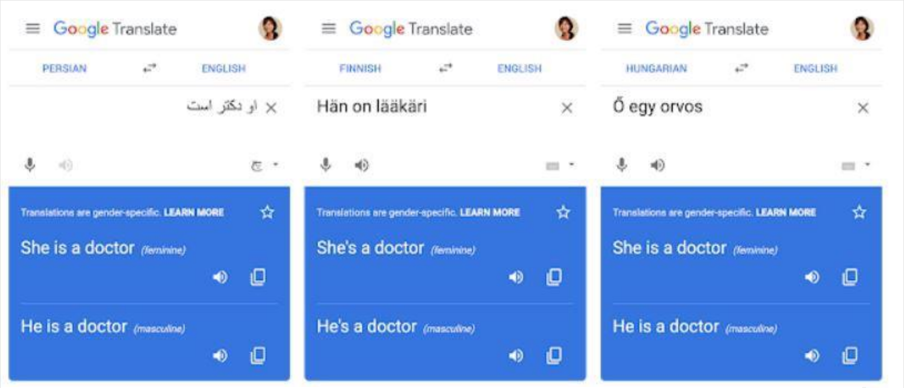


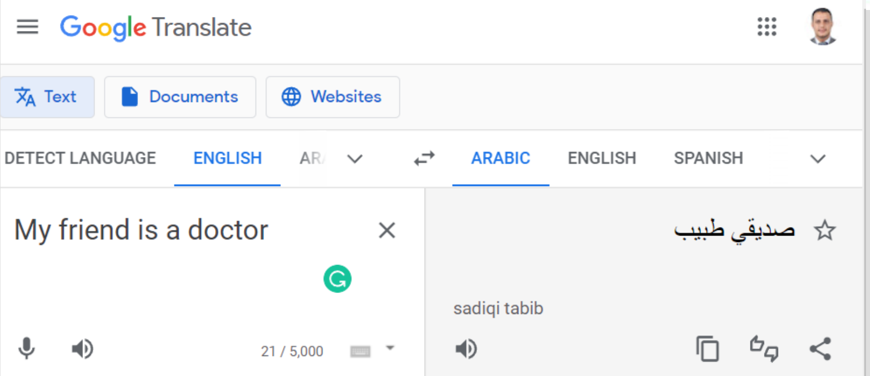
Bias in Interpretation

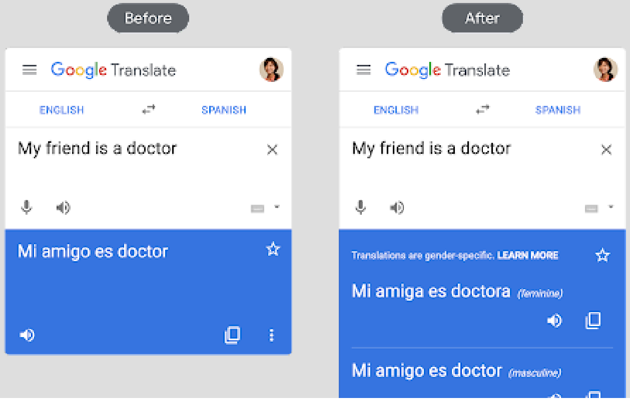


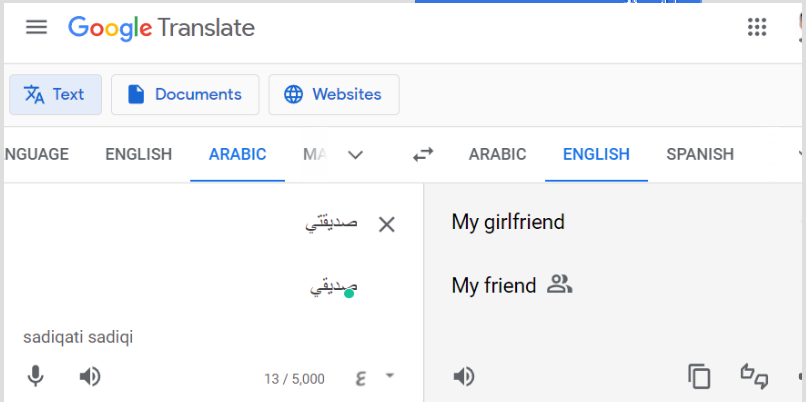
Biased statement
All housekeepers are women

How can we reduce gender bias in NLP systems

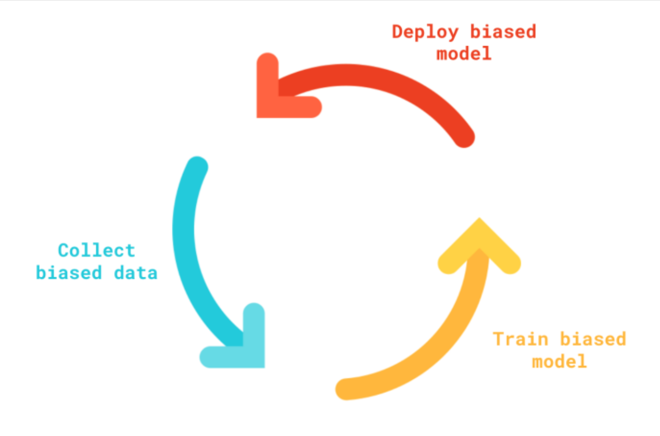








Human data perpetuates human biases NLP learns from human data, the results is a biased loo



How to mitigate gender/racial bias in NLP?

– Evaluation of Gender Bias in Word Embeddings

Vector Embedding of Words

- A word is represented as a vector.
- Word embeddings depend on a notion of ***word similarity***.

Similarity is computed using cosine.
- A very useful definition is paradigmatic similarity:

Similar words occur in ***similar contexts.*** They are ***exchangeable.***

POTUS
Yesterday The President called a press conference.
Trump

“POTUS: President of the United States.”

Vector Embedding of Words

– Traditional Method - Bag of Words Model

– Either uses one hot encoding.

Each word in the vocabulary is represented by one bit position in a HUGE vector.
For example, if we have a vocabulary of 10000 words, and "Hello" is the 4th word in the dictionary, it would be represented by: 0 0 0 1 0 0 . . . . 0 0 0

– Or uses document representation.

Each word in the vocabulary is represented by its presence in documents.
For example, if we have a corpus of 1M documents, and "Hello" is in 1th, 3th and 5th documents ***only***, it would be represented by: 1 0 1 0 1 0 . . . . 0 0 0

– Context information is not utilized.

– Word Embeddings

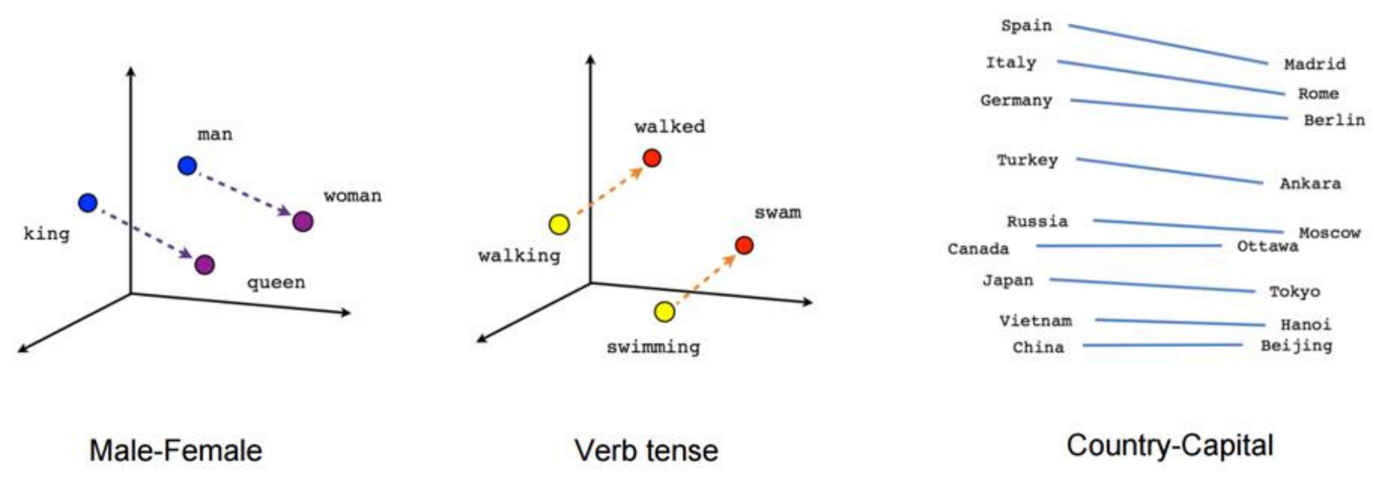
– Stores each word in as a point in space, where it is represented by a dense vector of fixed number of dimensions (generally 300) .

– Unsupervised, built just by reading huge corpus.

– For example, "Hello" might be represented as : [0.4, -0.11, 0.55, 0.3 . . . 0.1, 0.02].

– Dimensions are basically projections along different axes, more of a mathematical concept.

Example



- vector[Queen] ≈ vector[King] - vector[Man] + vector[Woman]
- vector[Paris] ≈ vector[France] - vector[Italy] + vector[Rome]

This can be interpreted as “France is to Paris as Italy is to Rome”.

Working with vectors

- Finding the most similar words to $\overrightarrow{dog}$.
- Compute the similarity from word $\overrightarrow{dog}$ to all other words.
- This is a single matrix-vector product: $\mathrm{W} \cdot \overrightarrow{dog}$
  - W is the word embedding matrix of |V| rows and d columns.
  - Result is a |V| sized vector of similarities.
  - Take the indices of the k-highest values.

Working with vectors

- Similarity to a group of words
- “Find me words most similar to cat, dog and cow”.
- Calculate the pairwise similarities and sum them:

$$\mathrm{W} \cdot \overrightarrow{cat} + \mathrm{W} \cdot \overrightarrow{dog} + \mathrm{W} \cdot \overrightarrow{cow}$$

- Now find the indices of the highest values as before.
- Matrix-vector products are wasteful. Better option:

$$\mathrm{W} \cdot (\overrightarrow{cat} + \overrightarrow{dog} + \overrightarrow{cow})$$

Applications of Word Vectors

- Word Similarity
- Machine Translation
- Part-of-Speech and Named Entity Recognition
- Relation Extraction
- Sentiment Analysis
- Co-reference Resolution
  - Chaining entity mentions across multiple documents - can we find and unify the multiple contexts in which mentions occurs?
- Clustering
  - Words in the same class naturally occur in similar contexts, and this feature vector can directly be used with any conventional clustering algorithms (K-Means, agglomerative, etc). Human doesn’t have to waste time hand-picking useful word features to cluster on.
- Semantic Analysis of Documents
- Build word distributions for various topics, etc.

Vector Embedding of Words

- In this lecture will describe the Word2Vec:
  - Prediction-based model.
  - Consider occurrences of terms at context level.

word2Vec: Local contexts

- Instead of entire documents, ***Word2Vec*** uses words ***k*** positions away from each center word.
- These words are called **context words**.
- Example for ***k=3***:

“It was a bright cold day in April, and the clocks were striking”.
Center word: red (also called focus word).
Context words: blue (also called target words).

- Word2Vec considers all words as center words, and all their context words.

Word2Vec: Data generation (window size = 2)

Example:

d1 = “king brave man” , d2 = “queen beautiful women”

word	Word one hot encoding	neighbor	Neighbor one hot encoding
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]	women	[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]	women	[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]

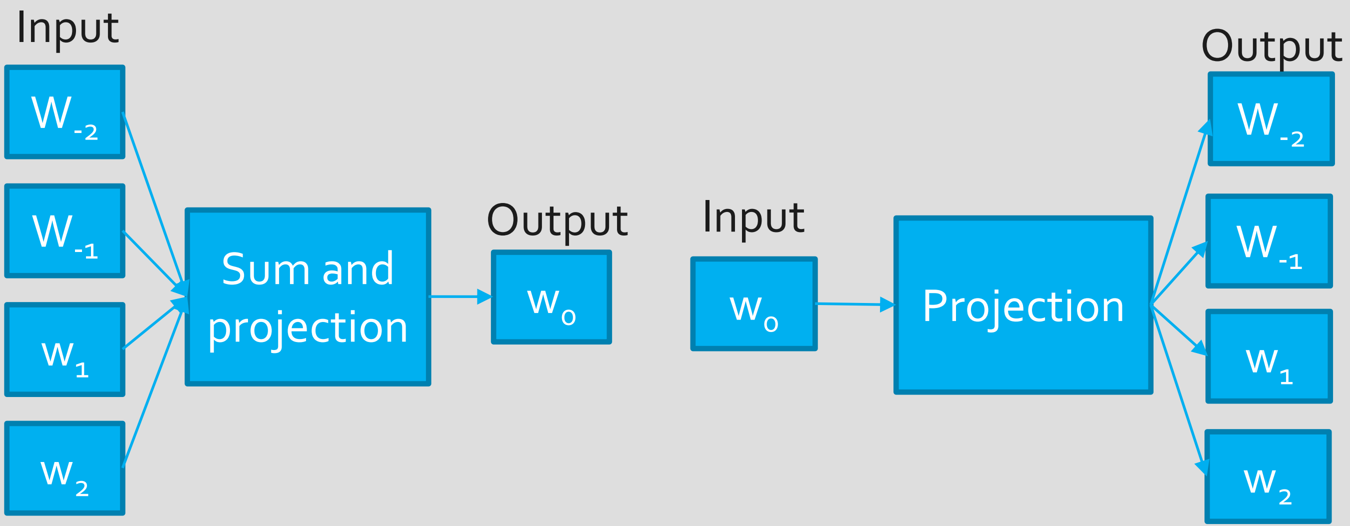
Word2Vec: Data generation (window size = 2)

Example:

d1 = “king brave man” , d2 = “queen beautiful women”

word	Word one hot encoding	neighbor	Neighbor one hot encoding
king	[1,0,0,0,0,0]	brave	[0,1,1,0,0,0]
		man	
brave	[0,1,0,0,0,0]	king	[1,0,1,0,0,0]
		man	
man	[0,0,1,0,0,0]	king	[1,1,0,0,0,0]
		brave	
queen	[0,0,0,1,0,0]	beautiful	[0,0,0,0,1,1]
		women	
beautiful	[0,0,0,0,1,0]	queen	[0,0,0,1,0,1]
		women	
woman	[0,0,0,0,0,1]	queen	[0,0,0,1,1,0]
		beautiful	

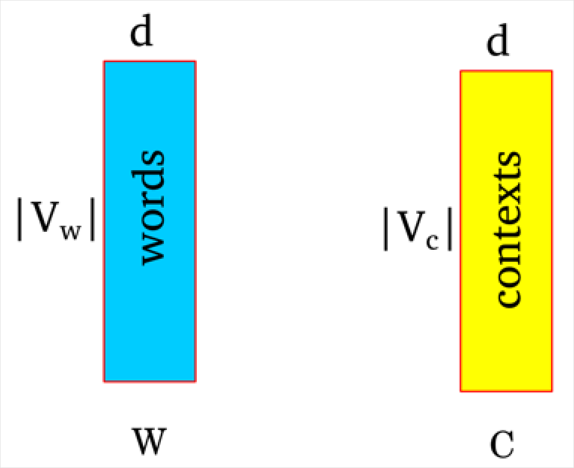
Word2Vec: main context representation models



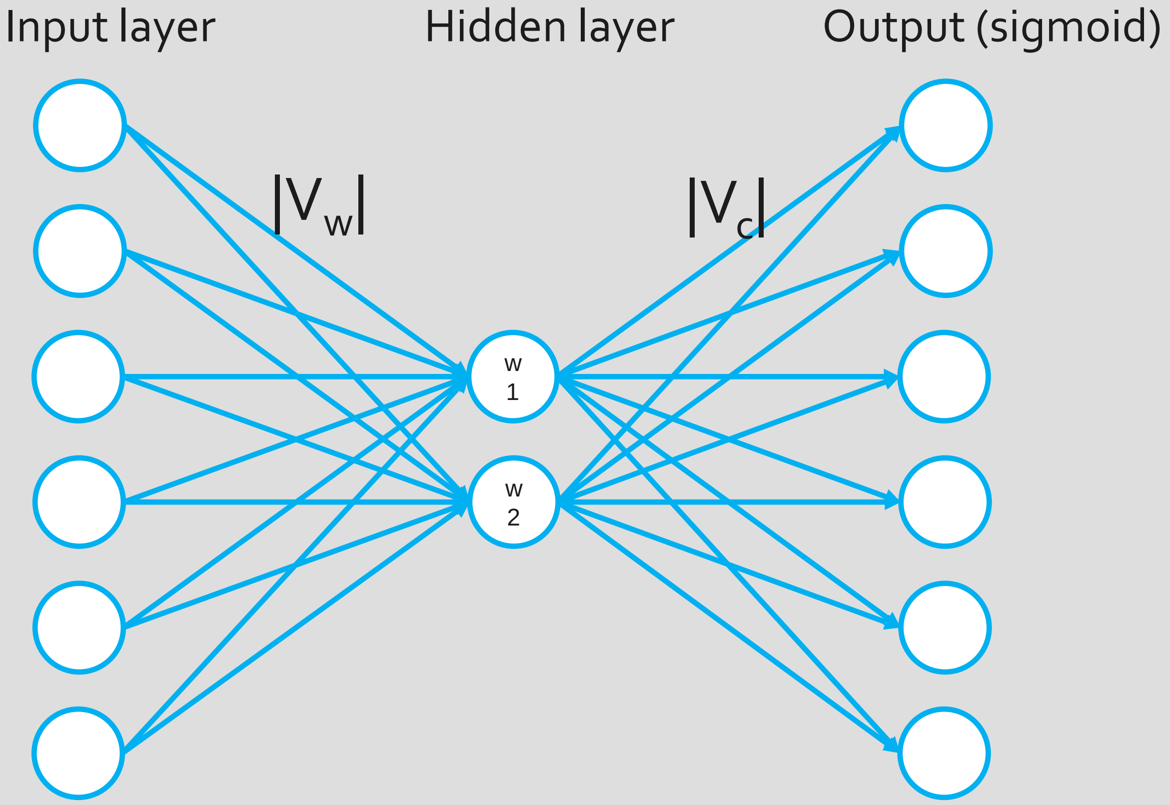
- Word2Vec is a predictive model.
- Will focus on Skip-Ngram model

How does word2Vec work?

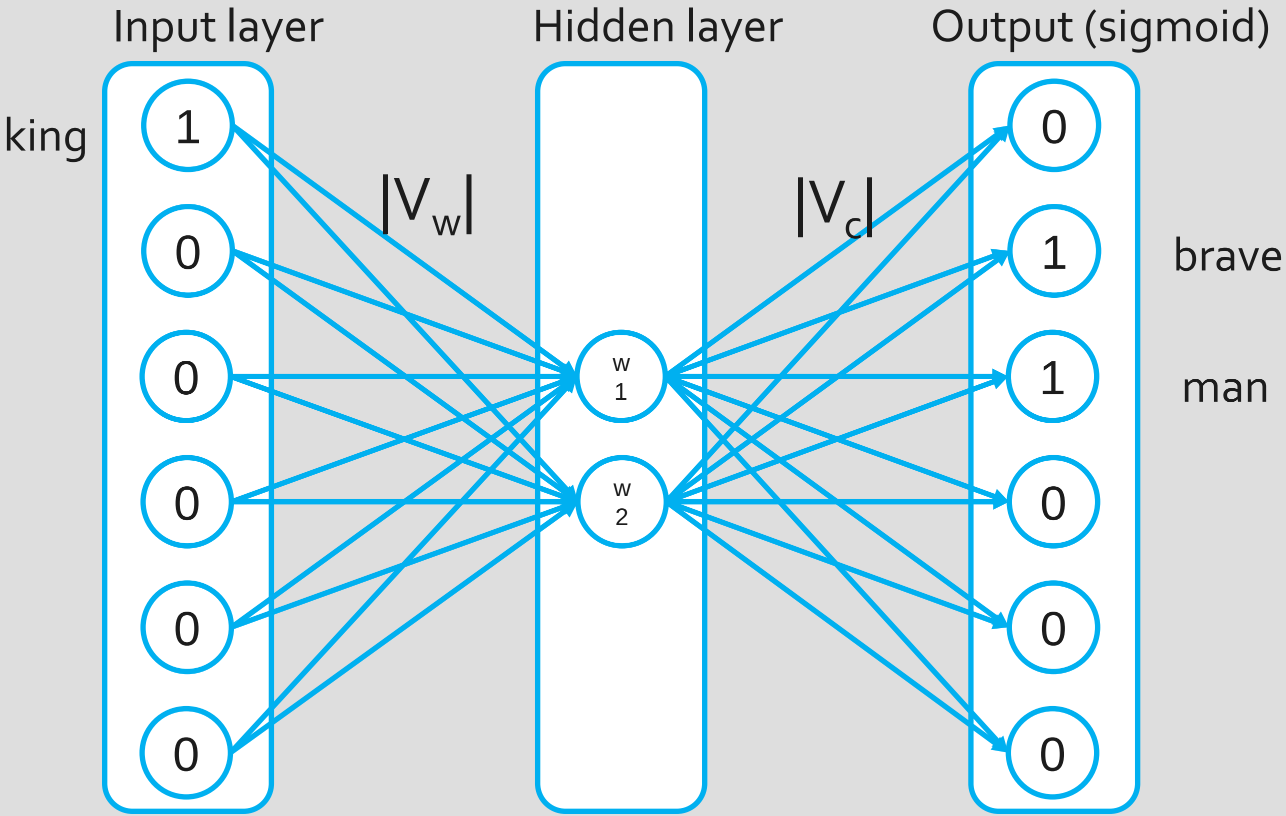
- Represent each word as a d dimensional vector.
- Represent each context as a d dimensional vector.
- Initialize all vectors to random weights.
- Arrange vectors in two matrices, W and C.



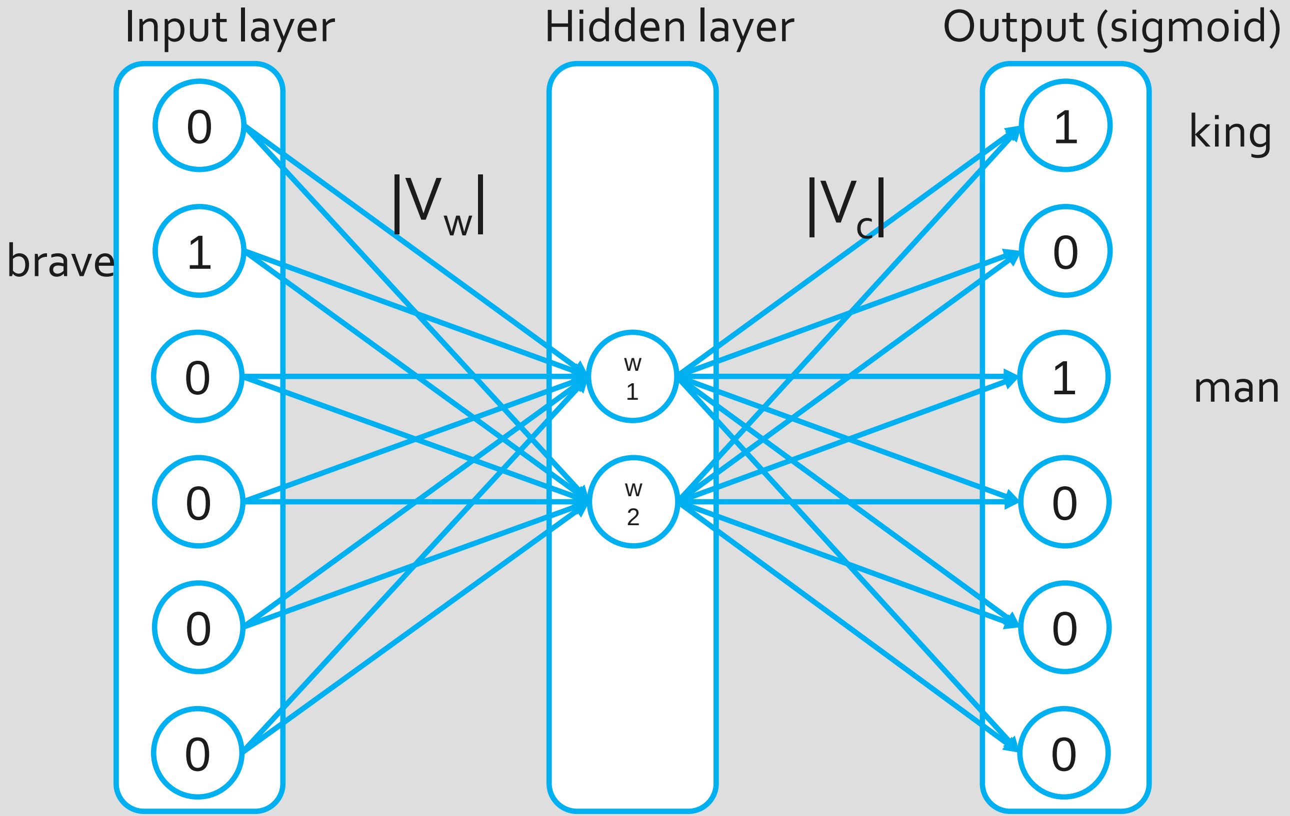
Word2Vec : Neural Network representation



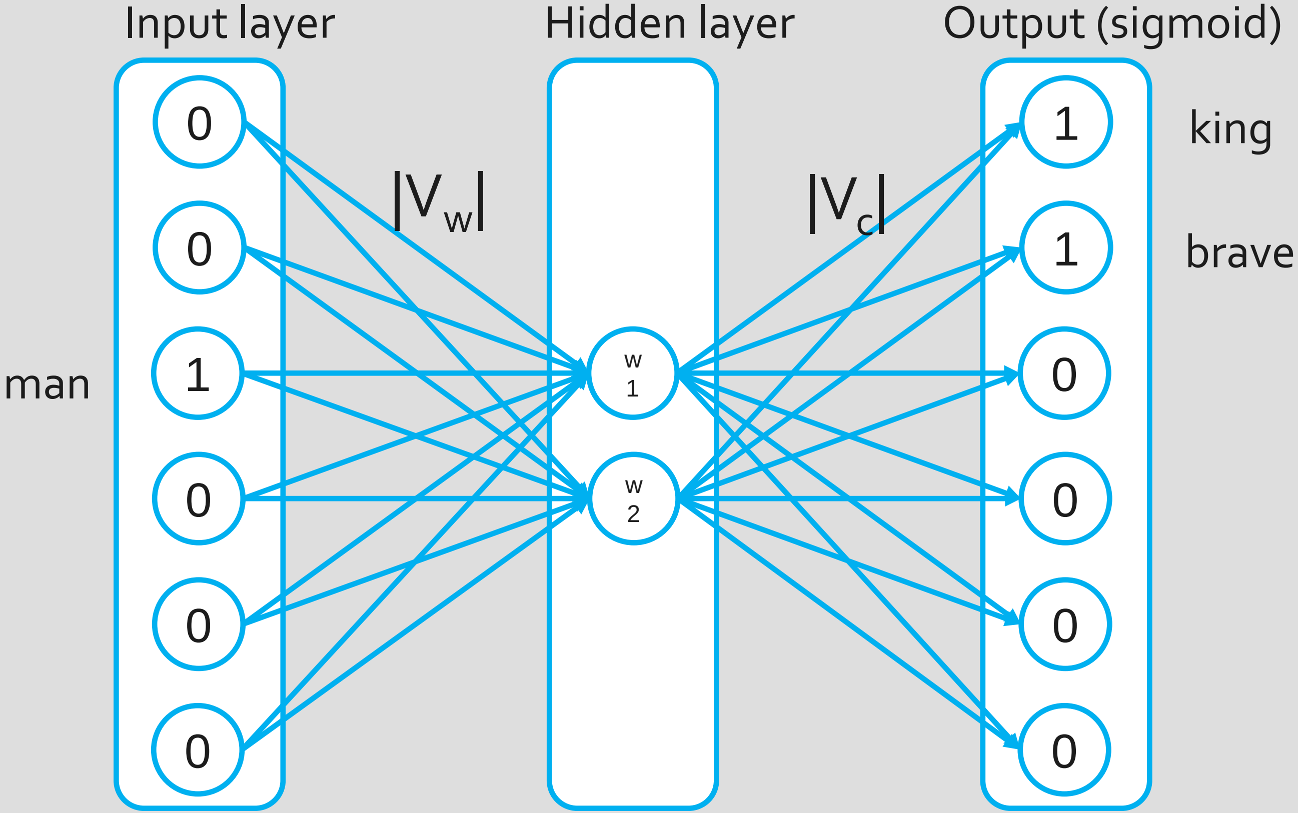
Word2Vec : Neural Network representation



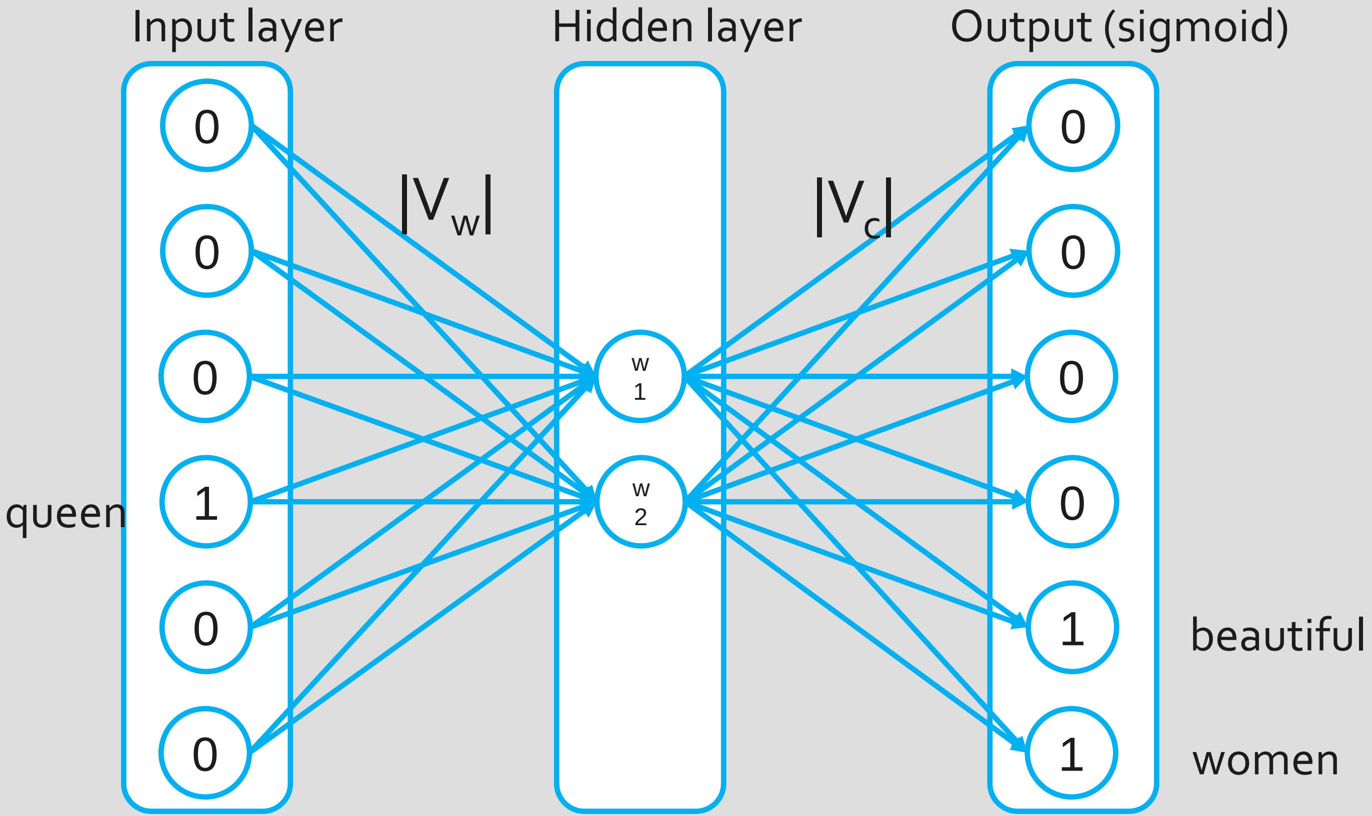
Word2Vec : Neural Network representation



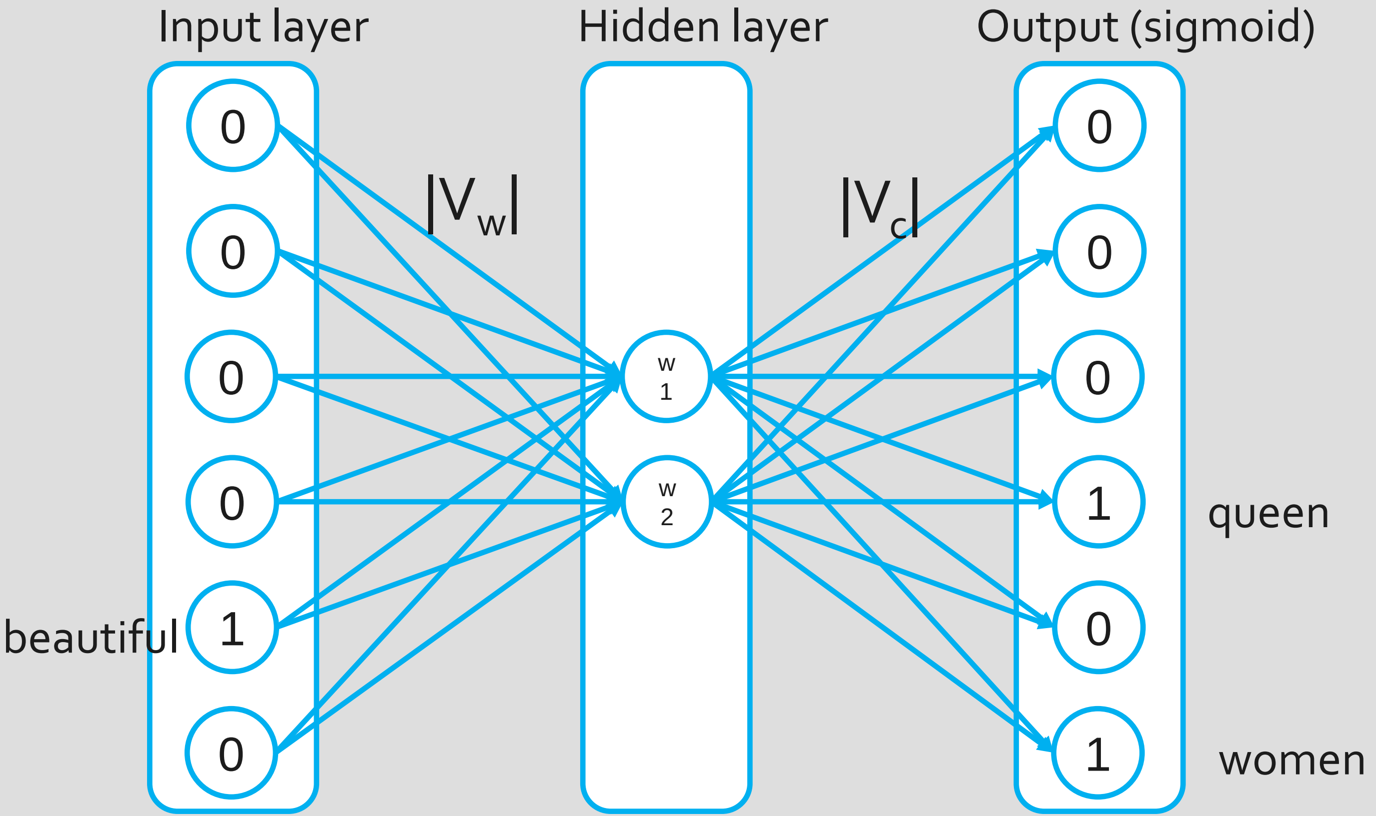
Word2Vec : Neural Network representation



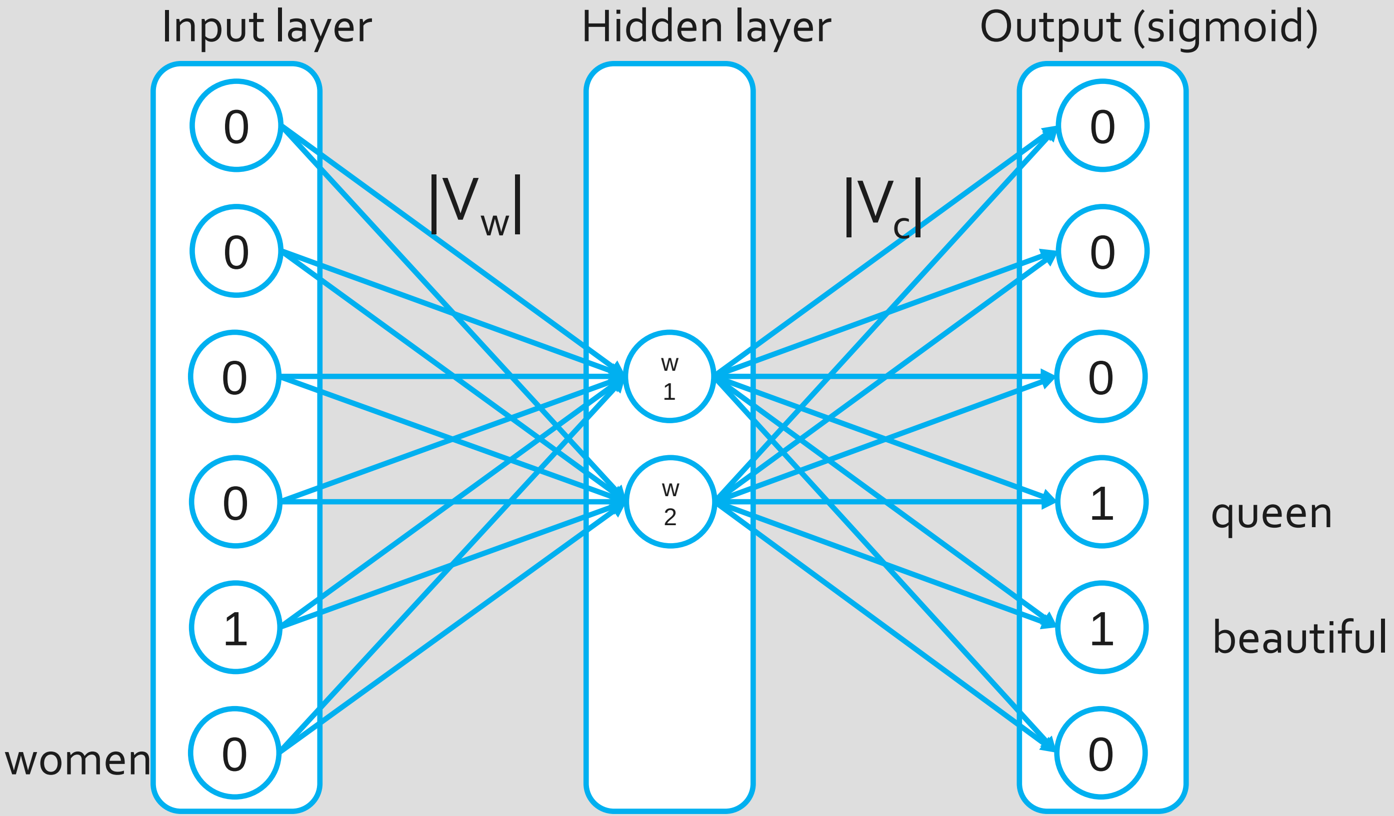
Word2Vec : Neural Network representation



Word2Vec : Neural Network representation



Word2Vec : Neural Network representation



Skip-Ngram: Training method

- The prediction problem is modeled using soft-max:

$$p(c|w;\theta) = \frac{\exp(v_c \cdot v_w)}{\sum_{\grave{c} \in C} \exp(v_{\grave{c}} \cdot v_w)}$$

Predict context words(s) ***c***
From focus word ***w***
Looks like logistic regression!
  - v_w are features and the evidence is v_c
- The objective function (in log space):

$$\underset{\theta}{\operatorname{argmax}} \sum_{(w,c) \in D} \log p(c|w;\theta) = \sum_{(w,c) \in D} \left[\log \exp(v_c \cdot v_w) - \log \sum_{\grave{c} \in C} \exp(v_{\grave{c}} \cdot v_w) \right]$$

Skip-Ngram: Negative sampling

– The objective function (in log space):

$$\underset{\theta}{\operatorname{argmax}} \sum_{(w,c)\in D} \log p(c|w;\theta) = \sum_{(w,c)\in D} \left[\log \exp(v_c \cdot v_w) - \log \sum_{\grave{c}\in C} \exp(v_{\grave{c}} \cdot v_w)\right]$$

While the objective function can be computed optimized, it is computationally expensive

- $p(c|w;\theta)$ is very expensive to compute due to the summation $\sum_{\grave{c}\in C} \exp(v_{\grave{c}} \cdot v_w)$

– Mikolov et al. proposed the negative-sampling approach as a more efficient way of deriving word embeddings:

$$\underset{\theta}{\operatorname{argmax}} \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c)\in \acute{D}} \log \sigma(-v_c \cdot v_w)$$

Skip-Ngram: Example

– While more text:

Extract a word window:

Try setting the vector values such that:

- $\sigma(w \cdot c_1) + \sigma(w \cdot c_2) + \sigma(w \cdot c_3) + \sigma(w \cdot c_4) + \sigma(w \cdot c_5) + \sigma(w \cdot c_6)$ is high!

Create a corrupt example by choosing a random word $\acute{w}$

Try setting the vector values such that:

- $\sigma(\acute{w} \cdot c_1) + \sigma(\acute{w} \cdot c_2) + \sigma(\acute{w} \cdot c_3) + \sigma(\acute{w} \cdot c_4) + \sigma(\acute{w} \cdot c_5) + \sigma(\acute{w} \cdot c_6)$ is low!

Skip-Ngram: How to select negative samples?

– Can sample using frequency.

Problem: will sample a lot of stop-words.

– Mikolov et al. proposed to sample using:

$$p(w_i) = \frac{f(w_i)^{3/4}}{\sum_j f(w_j)^{3/4}}$$

Not theoretically justified but works well in practice!

Relations Learned by Word2Vec

– A relation is defined by the vector displacement in the first column. For each start word in the other column, the closest displaced word is shown.

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

“Efficient Estimation of Word Representations in Vector Space” Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Arxiv 2013