Wszystkie tzw. modele tematyczne (*topic models*) bazują na tych samych dwóch istotnych założeniach:

Wszystkie tzw. modele tematyczne (*topic models*) bazują na tych samych dwóch istotnych założeniach:

- każdy **dokument** jest mieszaniną (złożeniem) różnych **tematów**,

Wszystkie tzw. modele tematyczne (*topic models*) bazują na tych samych dwóch istotnych założeniach:

- każdy **dokument** jest mieszaniną (złożeniem) różnych **tematów**,
- każdy **temat** stanowi mieszaninę **słów**

Wszystkie tzw. modele tematyczne (*topic models*) bazują na tych samych dwóch istotnych założeniach:

- każdy **dokument** jest mieszaniną (złożeniem) różnych **tematów**,
- każdy **temat** stanowi mieszaninę **słów**

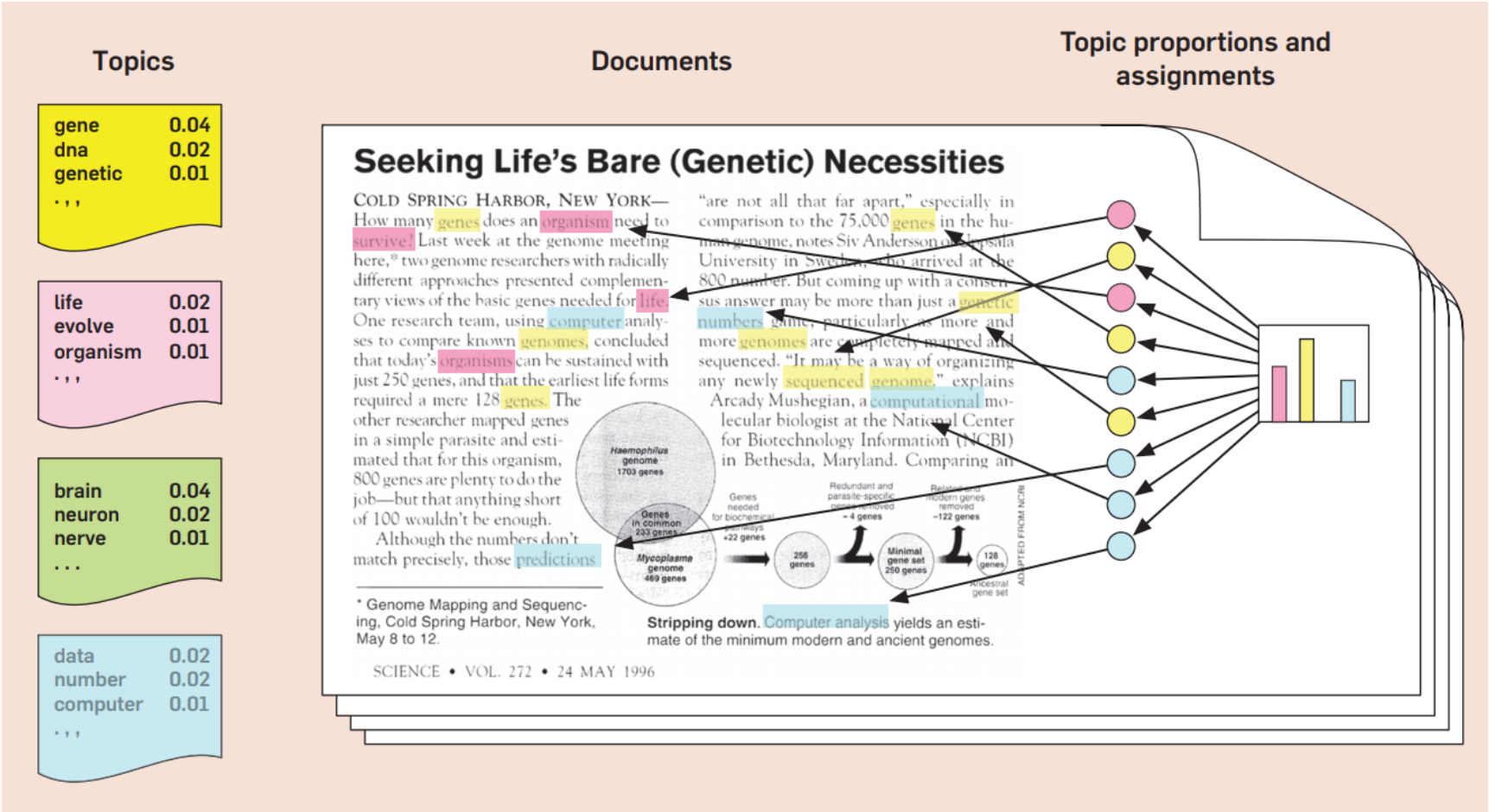
W praktyce oznacza to, że modele tematyczne budowane są wokół idei, mówiącej że ze strony semantycznej pojedynczy dokument jest tworzony za pomocą pewnych **ukrytych** czynników (lub składowych), których nie jesteśmy w stanie obserwować.

Wszystkie tzw. modele tematyczne (*topic models*) bazują na tych samych dwóch istotnych założeniach:

- każdy **dokument** jest mieszaniną (złożeniem) różnych **tematów**,
- każdy **temat** stanowi mieszaninę **słów**

W praktyce oznacza to, że modele tematyczne budowane są wokół idei, mówiącej że ze strony semantycznej pojedynczy dokument jest tworzony za pomocą pewnych **ukrytych** czynników (lub składowych), których nie jesteśmy w stanie obserwować.
W takim razie, widać już co jest głównym celem MT – odkrywanie (za pomocą różnych technik) tematów, które nadają kształt rozpatrywanym dokumentom.

Sens modelu tematycznego dobrze oddaje poniższy rysunek:



Dokument nr 1

If you want to cause a commotion in any psychology department or any other place where animal and human behaviour is studied, all that you have to do is to claim that your dog loves you. Skeptics, critics, and even some ardent supporters will pour out into the halls to argue the pros and cons of that statement.

Among the skeptics you will find the veterinarian Fred Metzge, of Pennsylvania State University, who claims that dogs probably don't feel love in the typical way humans do. Dogs make investments in human beings because it works for them. They have something to gain from putting so-called emotions out there.

Metzger believes that dogs 'love' us only as long as we continue to reward their behaviours with treats and attention. For most dog owners, however, there is little doubt that dogs can truly love people.

Dokument nr 2

Emotions guide our lives in a million ways. Whether we're inclined to hide and avoid or ponder and express them, most of us don't realize the extent to which they are driving our thoughts and behavior.

Exploring our emotions is a worthy endeavor for anyone hoping to know and develop themselves, build healthy relationships, and pursue what they want in life. Recent research has even suggested that emotional intelligence is more important than IQ, showing that it "predicts over 54% of the variation in success" in relationships, health, and quality of life.

Our emotions can offer us clues into who we are as well as how we've been affected by our history. Many of our actions are initiated by emotion, which leads to the natural question of what emotions are being surfaced and why.

Dokument nr 3

Curiosity is part of human nature. One of the first questions children learn to ask is “why?” As adults, we continue to wonder. Using empirical methods, psychologists apply that universal curiosity to collect and interpret research data to better understand and solve some of society’s most challenging problems.

It’s difficult, if not impossible, to think of a facet of life where psychology is not involved. Psychologists employ the scientific method — stating the question, offering a theory and then constructing rigorous laboratory or field experiments to test the hypothesis.

Psychologists apply the understanding gleaned through research to create evidence-based strategies that solve problems and improve lives.

Dokument nr 4

Olga, a 22-year-old woman in Saratov, Russia took her dog and her baby son Vadim to a park and met up with friends.

After a few drinks, Olga went home and left her baby behind! Luckily, her dog Lada was with the baby. Olga woke the next morning and realized the child was missing.

She thought Vadim had been abducted, but her father went to the park and found the baby in his pram, with Lada still beside him. The rottweiler had stood guard over him all night long. Vadim was wet and hungry, but unharmed, and was placed in the care of his grandmother.

Czyszczenie tekstów

W stosunku do każdego z dokumentów wykonujemy dobrze znane funkcje:

- usunięcie znaków interpunkcyjnych,
- usunięcie znaków specjalnych (np \n),
- usunięcie nadmiarowych spacji,
- przekształcenie na małe litery,
- usunięcie słów funkcyjnych (stopwords),
- stemowanie (za pomocą stemmera Portera)

Dokument 1

if want caus commot psycholog depart place anim human behaviour studi claim dog love skeptic critic even ardent support will pour hall argu pros con statement among skeptic will find veterinarian fred metzg pennsylvania state univers claim dog probabl feel love typic way human dog make invest human be work they someth gain put call emot metzger believ dog love us long continu reward behaviour treat attentionfor dog owner howev littl doubt dog can truli love peopl

Dokument 2

emot guid live million way whether re inclin hide avoid ponder express us dont realiz extent drive thought behavior explor emot worthi endeavor anyon hope know develop build healthi relationship pursu want life recent research even suggest emot intellig import iq show predict 54 variat success relationship health qualiti life our emot can offer us clue well ve affect histori mani action initi emot lead natur question emot surfac

Dokument 3

curios part human natur one first question children learn ask as adult continu wonder use empir method psychologist appli univers curios collect interpret research data better understand solv societi challeng problem it difficult imposs think facet life psycholog involv psychologist employ scientif method state question offer theori construct rigor laboratori field experi test hypothesi psychologist appli understand glean research creat evidence bas strategi solv problem improv live

Dokument 4

olga 22 yearold woman saratov russia took dog babi son vadim park met friend after drink olga went home left babi behind luckili dog lada babi olga woke next morn realiz child miss she thought vadim abduct father went park found babi pram lada still besid the rottweil stood guard night long vadim wet hungri unharm place care grandmoth

Zastosowanie tzw. Latent Dirichlet Allocation (LDA) do tych danych daje następujące wyniki:

Temat 1

- emot
- life
- psychologist
- question
- research
- appli
- curios

Temat 2

- babi
- olga
- vadim
- dog
- lada
- park
- went

Temat 3

- dog
- love
- human
- behaviour
- claim
- skeptic
- will

	Dok. 1	**Dok. 2**	**Dok. 3**	**Dok. 4**
1. temat	3	1	1	2
2. temat	2	2	3	3
3. temat	1	3	2	1

Dokument 1

if want caus commot psycholog depart place anim human behaviour studi claim dog love skeptic critic even ardent support will pour hall argu pros con statement among skeptic will find veterinarian fred metzg pennsylvania state univers claim dog probabl feel love typic way human dog make invest human be work they someth gain put call emot metzger believ dog love us long continu reward behaviour treat attentionfor dog owner howev littl doubt dog can truli love peopl

Dokument 2

emot guid live million way whether re inclin hide avoid ponder express us dont realiz extent drive thought behavior explor emot worthi endeavor anyon hope know develop build healthi relationship pursu want life recent research even suggest emot intellig import iq show predict 54 variat success relationship health qualiti life our emot can offer us clue well ve affect histori mani action initi emot lead natur question emot surfac

Dokument 3

curios part human natur one first question children learn ask as adult continu wonder use empir method psychologist appli univers curios collect interpret research data better understand solv societi challeng problem it difficult imposs think facet life psycholog involv psychologist employ scientif method state question offer theori construct rigor laboratori field experi test hypothesi psychologist appli understand glean research creat evidence bas strategi solv problem improv live

Dokument 4

olga 22 yearold woman saratov russia took dog babi son vadim park met friend after drink olga went home left babi behind luckili dog lada babi olga woke next morn realiz child miss she thought vadim abduct father went park found babi pram lada still besid the rottweil stood guard night long vadim wet hungri unharm place care grandmoth

LSA - Latent Semantic Analysis

LSA

- jedna z podstawowych prac z dziedziny,

LSA - Latent Semantic Analysis

LSA

- jedna z podstawowych prac z dziedziny,
- autorzy wystąpili nawet (i otrzymali) patent w USA (1988),

LSA - Latent Semantic Analysis

LSA

- jedna z podstawowych prac z dziedziny,
- autorzy wystąpili nawet (i otrzymali) patent w USA (1988),
- bazuje na znanej nam macierzy dokumentów-termów **A**,

LSA - Latent Semantic Analysis

LSA

- jedna z podstawowych prac z dziedziny,
- autorzy wystąpili nawet (i otrzymali) patent w USA (1988),
- bazuje na znanej nam macierzy dokumentów-termów **A**,
- najczęściej macierz ta jest wypełniana wartościami **tf-idf** zmiast zwyklych zliczen słów,

LSA - Latent Semantic Analysis

LSA

- jedna z podstawowych prac z dziedziny,
- autorzy wystąpili nawet (i otrzymali) patent w USA (1988),
- bazuje na znanej nam macierzy dokumentów-termów **A**,
- najczęściej macierz ta jest wypełniana wartościami **tf-idf** zmiast zwyklych zliczen słów,
- zakładamy, że chcemy otrzymać z tej analizy dwie kluczowe macierze: **macierz dokumentów-tematów** oraz **macierz tematów-słów**,

LSA - Latent Semantic Analysis

LSA

- jedna z podstawowych prac z dziedziny,
- autorzy wystąpili nawet (i otrzymali) patent w USA (1988),
- bazuje na znanej nam macierzy dokumentów-termów **A**,
- najczęściej macierz ta jest wypełniana wartościami **tf-idf** zmiast zwyklych zliczen słów,
- zakładamy, że chcemy otrzymać z tej analizy dwie kluczowe macierze: **macierz dokumentów-tematów** oraz **macierz tematów-słów**,
- głównym problemem jest fakt, że macierz **A** jest bardzo dużą i bardzo rzadką macierzą

LSA

- aby obejść ten problem wykorzystuje się metodę przyciętego rozkładu według wartości osobliwych (*truncated Singular Value Decomposition*).,

LSA

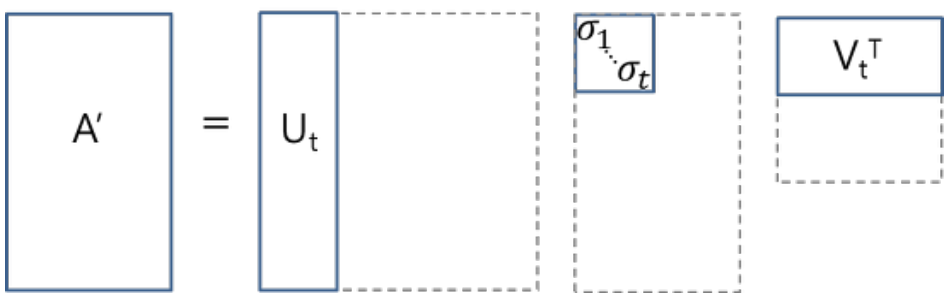
- aby obejść ten problem wykorzystuje się metodę przyciętego rozkładu według wartości osobliwych (*truncated Singular Value Decomposition*).,
- samo SVD umożliwia po prostu rozkład danej macierzy **A** na trzy macierze $\mathbf{A} = \mathbf{USV}^T$,

LSA

- aby obejść ten problem wykorzystuje się metodę przyciętego rozkładu według wartości osobliwych (*truncated Singular Value Decomposition*).,
- samo SVD umożliwia po prostu rozkład danej macierzy **A** na trzy macierze $\mathbf{A} = \mathbf{USV}^T$,
- w naszym przypadku ograniczamy się jedynie do t wartości osobliwych macierzy **S**,

LSA

- aby obejść ten problem wykorzystuje się metodę przyciętego rozkładu według wartości osobliwych (*truncated Singular Value Decomposition*).,
- samo SVD umożliwia po prostu rozkład danej macierzy **A** na trzy macierze $\mathbf{A} = \mathbf{USV}^T$,
- w naszym przypadku ograniczamy się jedynie do t wartości osobliwych macierzy **S**,



U jest macierzą dokumentów-tematów (kolumny to tematy), natomiast **V** to macierz słów-tematów (znów kolumny to tematy)

W przypadku **pLSA - Probabilistic Latent Semantic Analysis**, ten sam problem jest rozpatrywany z **probabilistycznego** punktu widzenia.

LSA

- chcemy znaleźć probabilistyczny model z ukrytymi tematami, który umożliwa **tworzenie** obserwowanej macierzy DTM,

W przypadku **pLSA - Probabilistic Latent Semantic Analysis**, ten sam problem jest rozpatrywany z **probabilistycznego** punktu widzenia.

LSA

- chcemy znaleźć probabilistyczny model z ukrytymi tematami, który umożliwa **tworzenie** obserwowanej macierzy DTM,
- zakładamy, że mając dokument d, temat z jest obecny w dokumencie z prawdopodobieństwem $P(z|d)$,

W przypadku **pLSA - Probabilistic Latent Semantic Analysis**, ten sam problem jest rozpatrywany z **probabilistycznego** punktu widzenia.

LSA

- chcemy znaleźć probabilistyczny model z ukrytymi tematami, który umożliwa **tworzenie** obserwowanej macierzy DTM,
- zakładamy, że mając dokument d, temat z jest obecny w dokumencie z prawdopodobieństwem $P(z|d)$,
- z drugiej strony dla danego tematu z, słowo w jest losowane z z w prawdopodobieństwem $P(w|z)$,

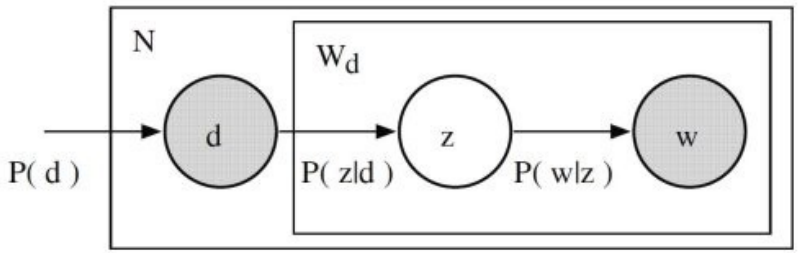
W przypadku **pLSA - Probabilistic Latent Semantic Analysis**, ten sam problem jest rozpatrywany z **probabilistycznego** punktu widzenia.

LSA

- chcemy znaleźć probabilistyczny model z ukrytymi tematami, który umożliwa **tworzenie** obserwowanej macierzy DTM,
- zakładamy, że mając dokument d, temat z jest obecny w dokumencie z prawdopodobieństwem $P(z|d)$,
- z drugiej strony dla danego tematu z, słowo w jest losowane z z w prawdopodobieństwem $P(w|z)$,
- czyli formalnie łączne prawdopodobieństwo obserwacji danego dokumentu i słowa razem $P(d, w)$ jest dane jako:

$$P(d, w) = P(d) \sum_z P(z|d) P(w|z)$$

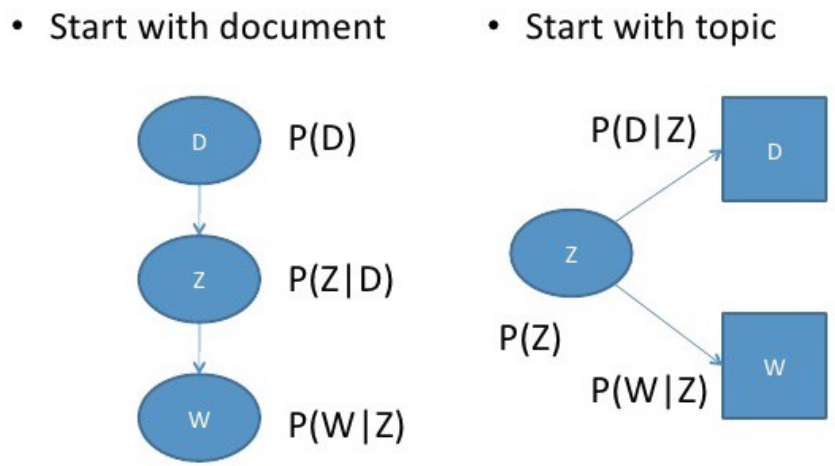
Taki schemat można również przedstawić jako:



Co istotne $P(w, d)$ można również przedstawić w inny sposób, mianowicie

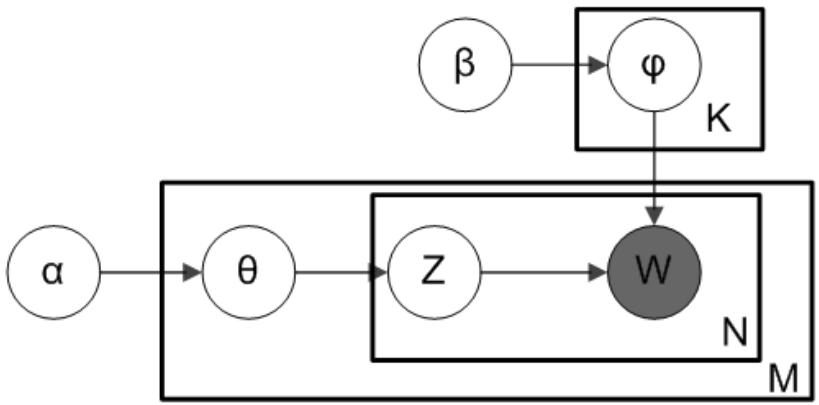
$$P(d, w) = \sum_z P(z)P(d|z)P(w|z)$$

czyli startujemy tym razem nie z dokumentu, tylko z rozkładu tematów



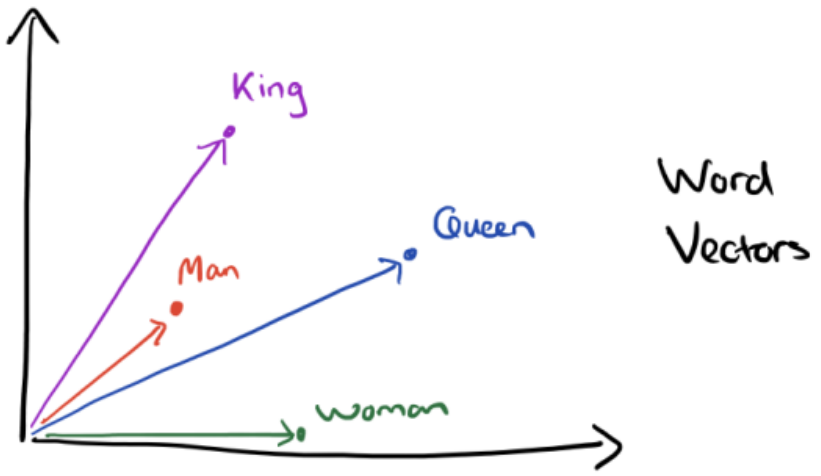
LDA, czyli **Latent Dirichlet Allocation** jest Bayesowska wersja pLSA.

W pLSA próbkowaliśmy dokument, a następnie temat bazując na dokumencie i w końcu słowa bazując na temacie. Dla LDA schemat wygląda następująco

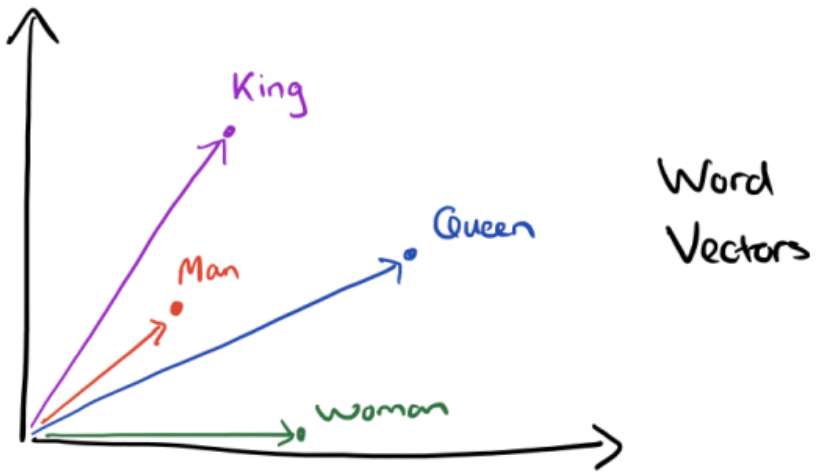


- tu najpierw próbkujemy **rozkład tematów** θ z pewnego rozkładu, zwanego rozkładem Dirichleta parametryzowanego przez wartość α
- z rozkładu θ losujemy konkretny temat *z*
- potem próbkujemy **rozkład słów** ϕ z pewnego rozkładu, zwanego rozkładem Dirichleta parametryzowanego przez wartość β
- z rozkładu ϕ losujemy konkretne słowo *w*

Jednym z częstych zagadnień związanych z text mining jest budowa tzw. **word embedding**, czyli sposobu reprezentacji wyrazów w postaci liczbowej. W pewien sposób mówiliśmy już o tym, kiedy rozpatrywaliśmy **model przestrzeni wektorowej** - jest to jednak bardzo ograniczona reprezentacja.



Jednym z częstych zagadnień związanych z text mining jest budowa tzw. **word embedding**, czyli sposobu reprezentacji wyrazów w postaci liczbowej. W pewien sposób mówiliśmy już o tym, kiedy rozpatrywaliśmy **model przestrzeni wektorowej** - jest to jednak bardzo ograniczona reprezentacja.



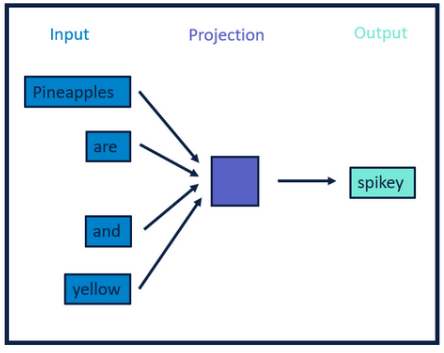
... and the *cute* **kitten** *purred* and then ...
... the *cute furry* **cat** *purred* and *miaowed* ...
... that the *small* **kitten** *miaowed* and she ...
... the *loud furry* **dog** *ran* and *bit* ...

Example **basis vocabulary**: *{bit, cute, furry, loud, miaowed, purred, ran, small}*.

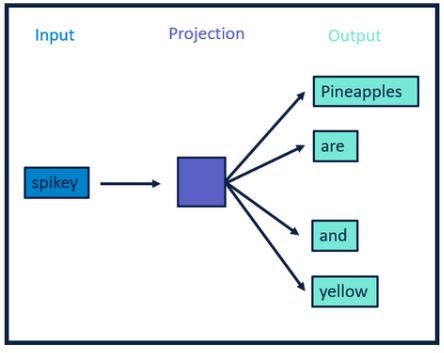
kitten context words: {cute, purred, small, miaowed}.
cat context words: {cute, furry, miaowed}.
dog context words: {loud, furry, ran, bit}.

Semantyka dystrybucyjna ma szerokie zastosowanie w rozwiązywaniu szeregu zadań związanych z przetwarzaniem języka naturalnego. U jej podstaw leży hipoteza, że słowa występujące w **podobnych kontekstach** w dużych zbiorach danych tekstowych mają **podobne znaczenie**.

Jednym z bardziej popularnych przykładów, należącym do tej grupy metod jest **word2vec**, stworzony przez Tomasa Mikolova z Google ok. 5 lat temu. Podejście to może wykorzystywać dwie różne techniki: **CBOW** - Continuous Bag of Words oraz **Skip-Gram**. Teoretycznie metody są algorytmicznie identyczne, oprócz tego, że CBOW przewiduje kluczowe słowo na podstawie kontekstu a Skip-Gram odwrotnie.

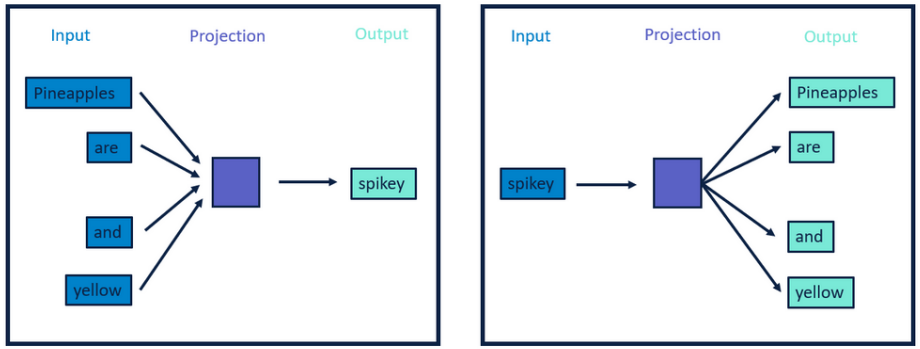


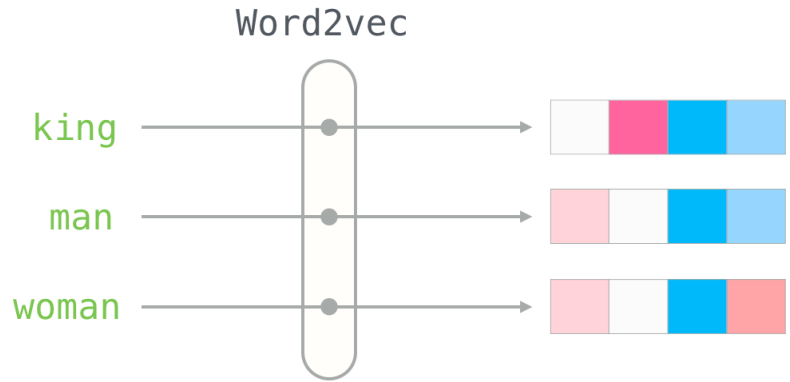
CBOW



Skip-gram

Jednym z bardziej popularnych przykładów, należącym do tej grupy metod jest **word2vec**, stworzony przez Tomasa Mikolova z Google ok. 5 lat temu. Podejście to może wykorzystywać dwie różne techniki: **CBOW** - Continuous Bag of Words oraz **Skip-Gram**. Teoretycznie metody są algorytmicznie identyczne, oprócz tego, że CBOW przewiduje kluczowe słowo na podstawie kontekstu a Skip-Gram odwrotnie.





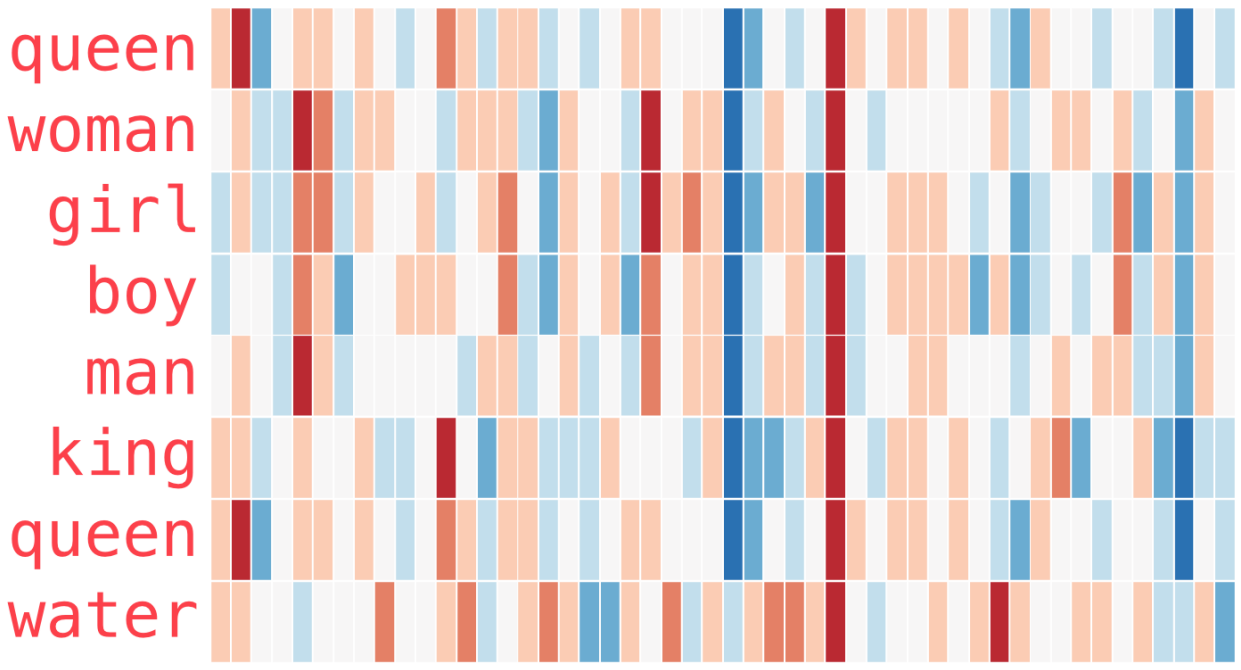
Word2vec używa sieci neuronowej z pojedynczą warstwą jako podstawowej architektury, ale jako ogólny wskaźnik używane jest po prostu prawdopodobieństwo $p(w_t|h)$ słowa w_t pod warunkiem historii (otoczenia) h, które jest wyznaczane przy użyciu metody największej wiarygodności.

GloVe, czyli **Global Vectors for Word Representation** jest datującą się na rok 2014 techniką zaproponowaną przez naukowców ze Stanford University i opierającą się na następującym algorytmie:

- wykonaj statystykę współwystępowania słów i zapisz ją w postaci macierzy $\mathbf{X}$; zwykle korpus jest skanowany w taki sposób, że szukamy słów kontekstowych w pewnym oknie zarówno przed jak i po interesującym nas słowie; zwykle też dalsze słowa wchodzą z mniejszą wagą np. $w = 1/przesuniecie$
- zdefiniuj więzy dla każdej pary słów występującej w macierzy $\mathbf{X}$ jako $\mathbf{w}_i^T \mathbf{w}_j + b_i + b = \log(\mathbf{X}_{ij})$, gdzie $\mathbf{w}_i$ to wektor głównego słowa, a $\mathbf{w}_j$ to wektor słowa z kontekstu, natomiast b_i, b to skalary
- zdefiniuj funkcję kosztu $J = \sum_i \sum_j (\mathbf{w}_i^T \mathbf{w}_j + b_i + b - \log(\mathbf{X}_{ij}))^2$

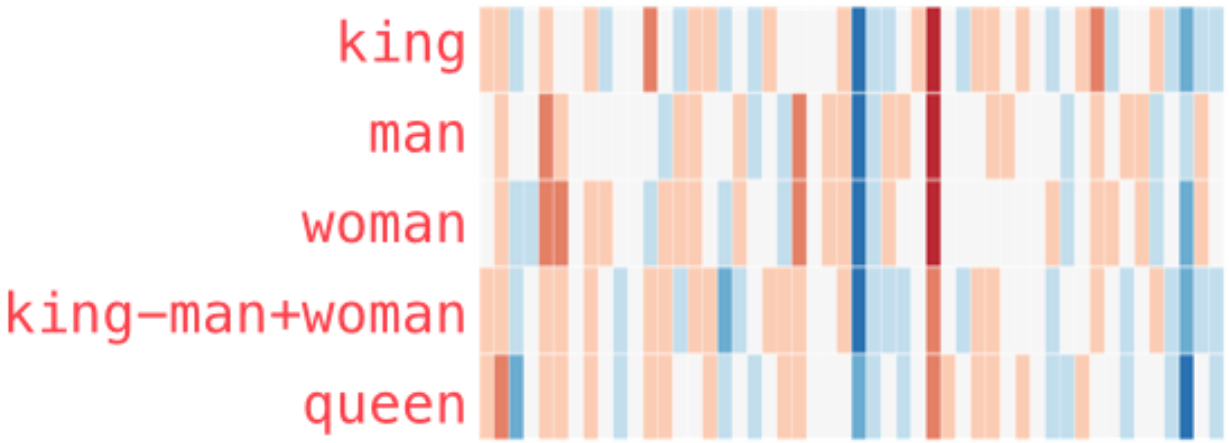
przy czym $f(\mathbf{X}_{ij})$ jest pewną funkcją ważącą, zapobiegającą uczeniu się jedynie od najczęściej wystepujących słów. Funkcja zostało zaproponowana jako $f(\mathbf{X}_{ij}) = (\mathbf{X}_{ij}/x_{max})^{\alpha}$ dla $\mathbf{X}_{ij} < x_{max}$ i $f(\mathbf{X}_{ij}) = 1$ w przeciwnym razie.

OK, ale co faktycznie z tego otrzymamy? Otóż dzięki powyższym zabiegom możemy uzyskać reprezentację słowa w pewnej przestrzeni (wymiary tej przestrzeni nie są związane z żadnymi konkretnymi słowami).

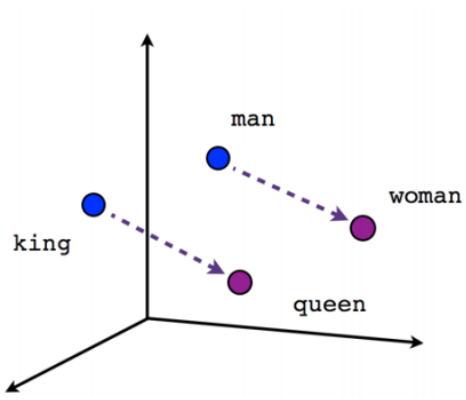


Co więcej, okazuje się, że w takiej przestrzeni dobrze działa wektorowa arytmetyka “semantyczna”.

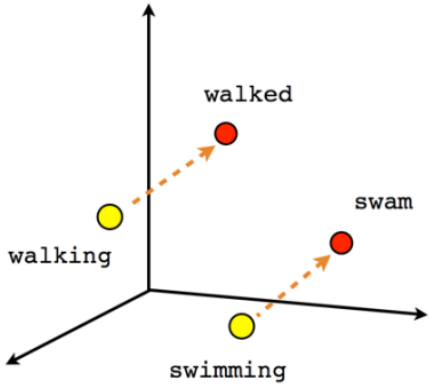
king − man + woman ~= queen



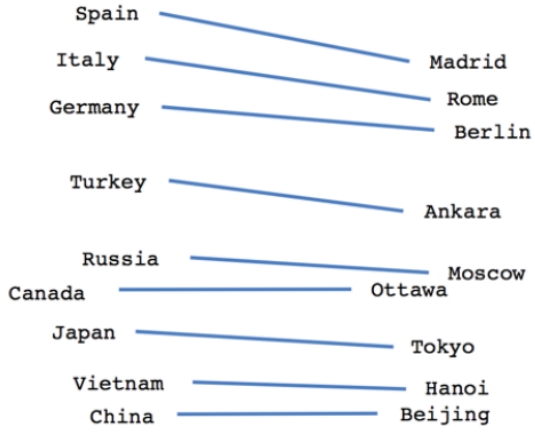
Jeśli natomiast dokonamy rzutowania za pomocą skalowania wielowymiarowego na przestrzeń dwuwymiarową, to można w łatwy sposób dostrzec analogie pomiędzy odpowiednimi parami słów (nie tylko rzeczowników).



Male-Female



Verb tense



Country-Capital