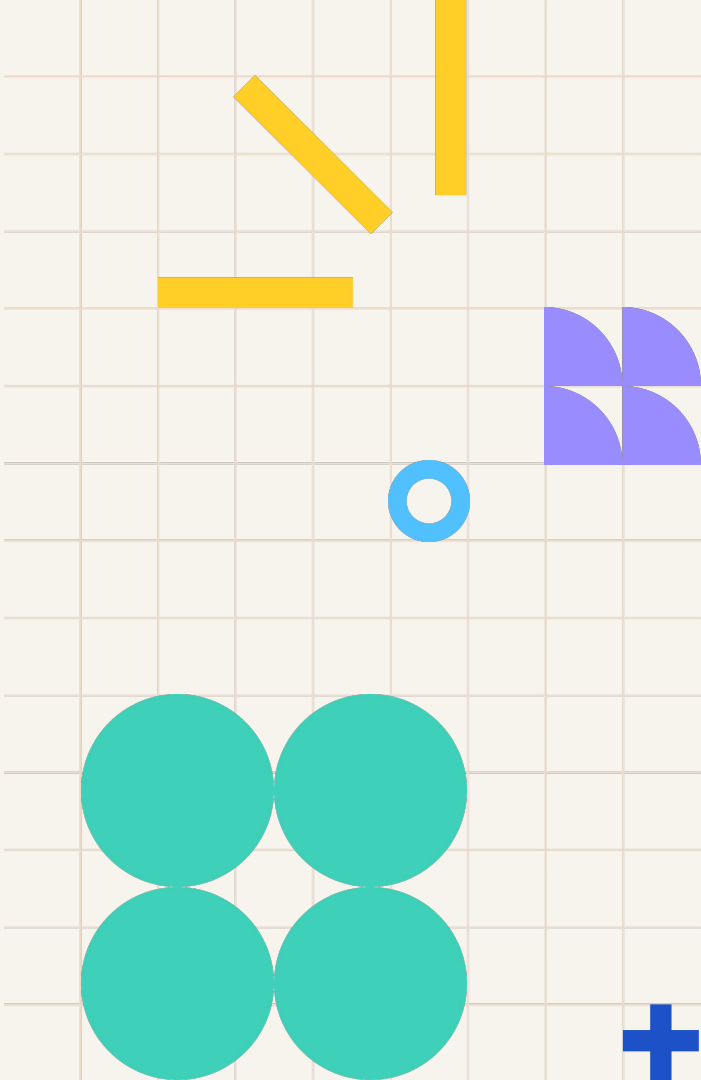
Introduction to NLP

Aydar Bulatov

DeepPavlov.ai
MIPT



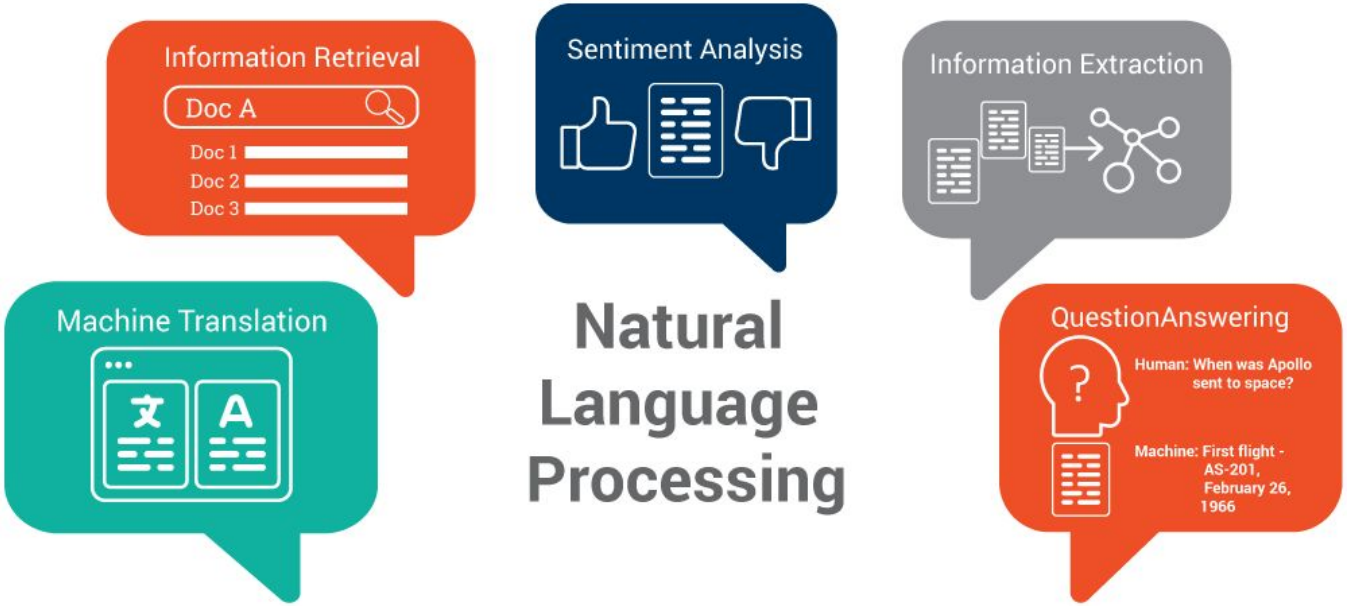
Today

01 What is NLP

02 Text representation

03 Text processing

What is Natural Language Processing?


Information Retrieval
Doc A
Doc 1
Doc 2
Doc 3
Sentiment Analysis
Information Extraction
Machine Translation
Natural Language Processing
QuestionAnswering
Human: When was Apollo sent to space?
Machine: First flight - AS-201, February 26, 1966

ML tasks in NLP

- Text classification
  - Sentiment, emotion, opinion analysis
  - Word sense disambiguation
- Sequence labelling
  - P-O-S tagging, named entity recognition, dialogue acts
- Parsing dependencies
- Coreference resolution

- Text generation
  - QA, summarization
  - data-to-text generation
  - dialogue systems
- Learning without supervision
  - clustering
  - matrix factorization
  - latent semantic indexing
- Semi-supervised learning
  - graph-based
  - clustering + classification

Word representations

Vector word representations

- dog [0.033, -0.009, 0.037, 0.068, ..., 0.053]
- cat [0.067, 0.069, -0.015, -0.0249, ... , 0.0317]
- home [-0.0004, 0.0199, 0.03, -0.04, ... , -0.031]
- computer [-0.0016 , 0.0872, -0.0314, -0.0523, ... , 0.0228]
- love [-0.169, 0.0382, 0.0084, -0.0182, ... , -0.0209]

Vector word representations

- dog [0.033, -0.009, 0.037, 0.068, ..., 0.053]
- cat [0.067, 0.069, -0.015, -0.0249, ... , 0.0317]
- home [-0.0004, 0.0199, 0.03, -0.04, ... , -0.031]
- computer [-0.0016 , 0.0872, -0.0314, -0.0523, ... , 0.0228]
- love [-0.169, 0.0382, 0.0084, -0.0182, ... , -0.0209]

- Wikipedia
- Twitter
- НКРЯ - Национальный корпус русского языка (собрание русских текстов в электронной форме)
- Тайга - корпус русского языка (в основе, художественная литература)

Vector word representations

Why not just use one-hot encoding?

- dog: [1, 0, 0, 0, ..., 0]
- cat: [0, 1, 0, 0, ... , 0]
- home: [0, 0, 1, 0, ... , 0]
- computer: [0, 0, 0, 1, ... , 0]
- love: [0, 0, 0, 0, ... , 1]

Vector word representations

Why not just use one-hot encoding?

- dog: [1, 0, 0, 0, ..., 0]
- cat: [0, 1, 0, 0, ... , 0]
- home: [0, 0, 1, 0, ... , 0]
- computer: [0, 0, 0, 1, ... , 0]
- love: [0, 0, 0, 0, ... , 1]

These vectors know nothing about the word!

Vector word representations

What properties should a representation share with its word?

Vector word representations

What properties should a representation share with its word?

- run ~ running
- walk ~ walked
- to describe ~ descriptor

Vector word representations

What properties should a representation share with its word?

- run ~ running
- walk ~ walked
- to describe ~ descriptor

- big ~ huge
- to buy ~ to purchase
- pineapple ~ coconut

Distributional semantics

You shall know a word by the company it keeps

Firth, 1957

Oculist and eye-doctor . . . occur in almost the same *environments*.

If A and B have almost identical *environments* we say that they are *synonyms*

Zellig Harris (1954)

Distributional semantics

Do you know what the word “*tezgüino*” means?

The example is from Jacob Eisenstein's NLP notes; the tezgüino example originally appeared in Lin, 1998.

AIRI × Университет Сириус

Distributional semantics

Do you know what the word “*tezgüino*” means?

But what if we know the following contexts:

1. a bottle of *tezgüino* is on the table,
2. everybody likes *tezgüino*,
3. *tezgüino* makes you drunk,
4. we make *tezgüino* out of corn.

The example is from Jacob Eisenstein's NLP notes; the tezgüino example originally appeared in Lin, 1998.

Distributional semantics

Do you know what the word “*tezgüino*” means?

But what if we know the following contexts:

1. a bottle of *tezgüino* is on the table,
2. everybody likes *tezgüino*,
3. *tezgüino* makes you drunk,
4. we make *tezgüino* out of corn.

Now we know!

Tezgüino is a kind of alcoholic beverage made from corn!

The example is from Jacob Eisenstein's NLP notes; the tezgüino example originally appeared in Lin, 1998.

Distributional semantics

Do you know what the word “*tezgüino*” means?

But what if we know the following contexts:

1. a bottle of *tezgüino* is on the table,
2. everybody likes *tezgüino*,
3. *tezgüino* makes you drunk,
4. we make *tezgüino* out of corn.

	context			
	1.	2.	3.	4.
tezgüino	1	1	1	1
loud	0	0	0	0
motor oil	1	0	0	1
tortillas	0	1	0	1
wine	1	1	1	0

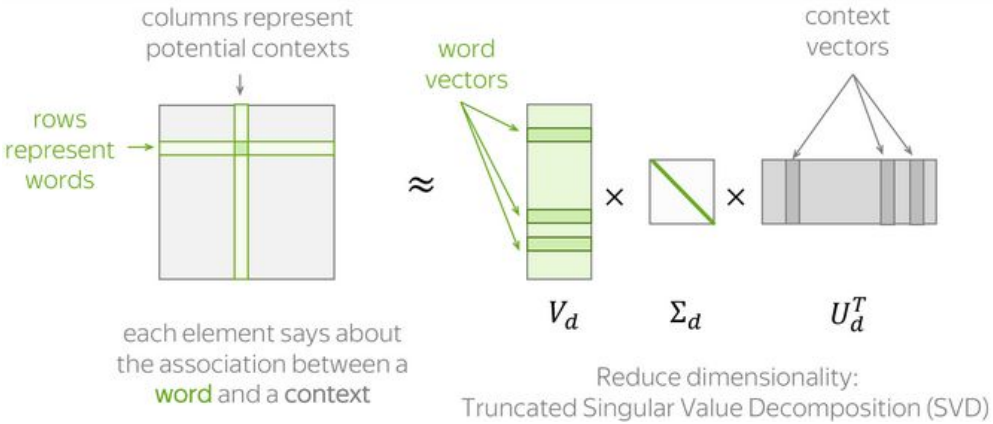
Now we know!

Tezgüino is a kind of alcoholic beverage made from corn!

The example is from Jacob Eisenstein's NLP notes; the tezgüino example originally appeared in Lin, 1998.

Count-based methods

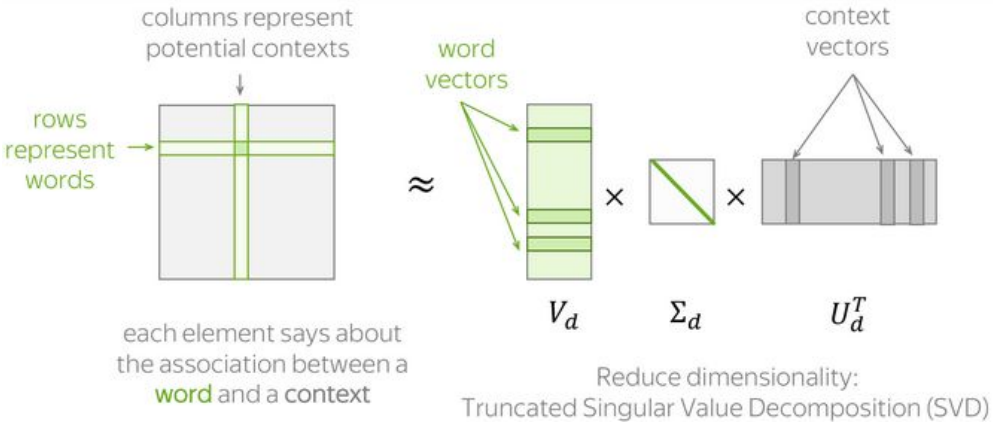
Main idea: We have to put information about contexts into word vectors.



images from Lena Voita NLP course

Count-based methods

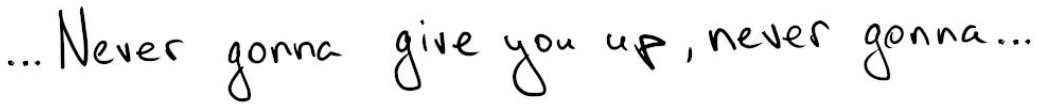
Main idea: We have to put information about contexts into word vectors.



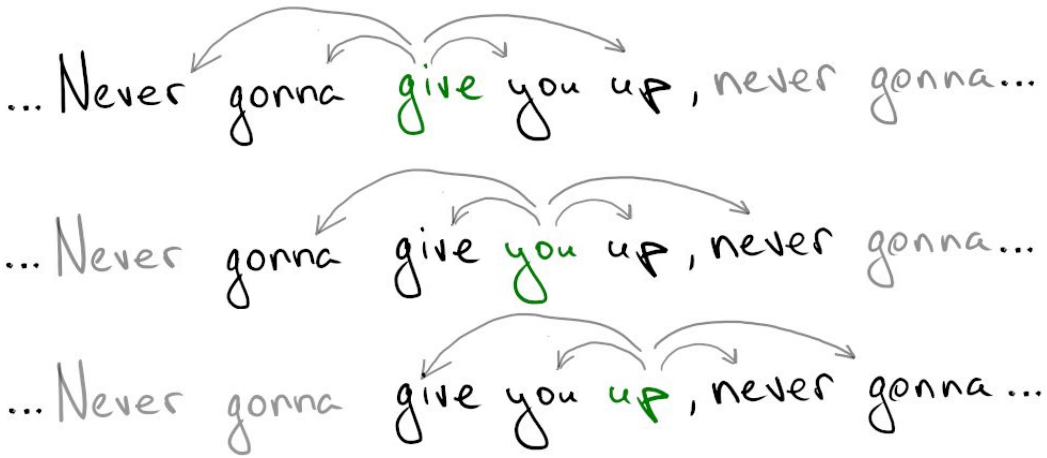
Ok but how do we represent context information?

images from Lena Voita NLP course

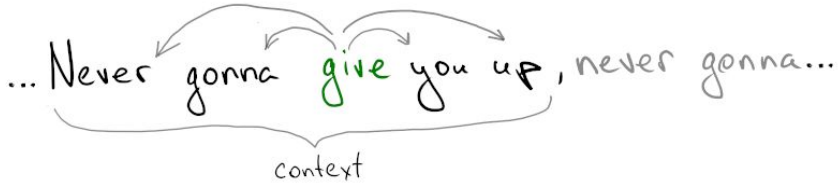
Count-based methods



Count-based methods



Count-based methods



1. Co-Occurrence counts

   $N(w, c)$ = number of times w appears in c

2. Tf-Idf

$$tf\text{-}idf(w, d, D) = tf(w, d) \cdot idf(w, D) = N(w, d) \cdot \log\left(\frac{|D|}{|\{d \in D : w \in D\}|}\right)$$

3. Positive PMI

   ppmi(w, c) = max(0, PMI(w, c))

$$pmi(w, c) = \log\left(\frac{P(w, c)}{P(w)\,P(c)}\right) = \log\left(\frac{N(w, c) \cdot |(w, c)|}{N(w)\,N(c)}\right)$$

Word2Vec

Idea: teach embeddings to predict their contexts.

Algorithm

1) use a large text corpus
2) for each word start with 2 vectors: central **u** + context **v**
3) go over the text with a sliding window (*use negative sampling*)
4) compute probabilities for words from c based on w
5) adjust the vectors to increase probabilities

$$P(o|c) = SM(u_o^T \cdot v_c) = \frac{e^{u_o^T v_c}}{\sum_{w \in V} e^{u_w^T v_c}}$$

Efficient Estimation of Word Representations in Vector Space (2013)

Word2Vec

Idea: teach embeddings to predict their contexts.

Algorithm

1) use a large text corpus
2) for each word start with 2 vectors: central **u** + context **v**
3) go over the text with a sliding window (*use negative sampling*)
4) compute probabilities for words from c based on w
5) adjust the vectors to increase probabilities

$$P(o|c) = SM(u_o^T \cdot v_c) = \frac{e^{u_o^T v_c}}{\sum_{w \in V} e^{u_w^T v_c}}$$

Objective function:

$$\text{Likelihood} = L(\theta) = \prod_{t=1}^{T} \prod_{-m \le j \le m, j \ne 0} P(w_{t+j}|w_t, \theta),$$

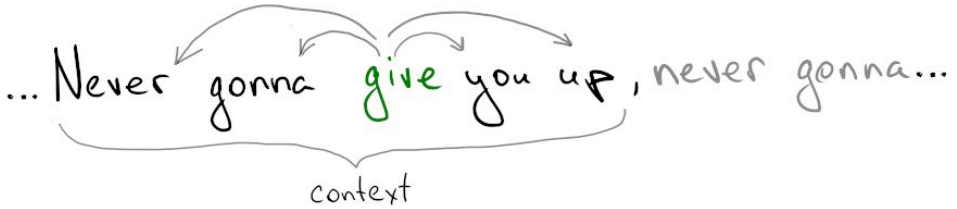
$$\text{Loss} = J(\theta) = -\frac{1}{T}\log L(\theta) = -\frac{1}{T}\sum_{t=1} \sum_{\substack{-m \le j \le m, \\ j \ne 0}} \log P(w_{t+j}|w_t, \theta)$$

agrees with our plan above ⟼ go over text; with a sliding window; compute probability of the context word given the central

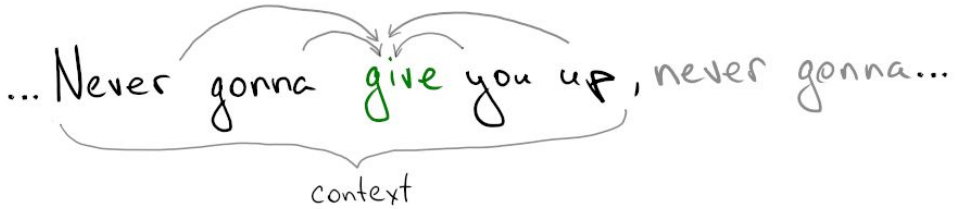
Efficient Estimation of Word Representations in Vector Space (2013)

Word2Vec

1. Skip-Gram
   from central word predict context



2. CBOW
   from sum of context predict word



Word2Vec


King
Queen
Woman
Princess
...
Royalty
Masculinity
Femininity
Age
...
0.99
0.99
0.05
0.7
0.99
0.05
0.93
0.6
0.02
0.01
0.999
0.5
0.98
0.02
0.94
0.1

GloVe

Idea: use **glo**bal corpus statistics to learn embedding **ve**ctors

W2V: $$J_{t,j}(\theta) = -\log P(cute|cat) = -\log \frac{\exp u_{cute}^T v_{cat}}{\sum_{w \in Voc} \exp u_w^T v_{cat}} = -u_{cute}^T v_{cat} + \log \sum_{w \in Voc} \exp u_w^T v_{cat}.$$

GloVe

Idea: use **glo**bal corpus statistics to learn embedding **ve**ctors

W2V: $$J_{t,j}(\theta) = -\log P(cute|cat) = -\log \frac{\exp u_{cute}^T v_{cat}}{\sum_{w \in Voc} \exp u_w^T v_{cat}} = -u_{cute}^T v_{cat} + \log \sum_{w \in Voc} \exp u_w^T v_{cat}.$$

GloVe: $$J(\theta) = \sum_{w,c \in V} f(\mathrm{N}(w,c)) \cdot (u_c^T v_w + b_c + \overline{b_w} - \log \mathrm{N}(w,c))^2$$

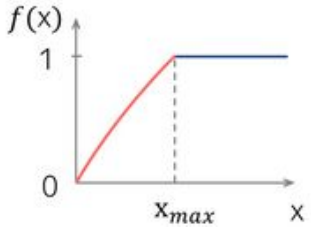
context vector

word vector

bias terms (also learned)

Weighting function to:

- penalize rare events
- not to over-weight frequent events



$$\begin{cases} (x/x_{max})^\alpha & \text{if } x < x_{max}, \\ 1 & \text{otherwise.} \end{cases}$$

$\alpha = 0.75, x_{max} = 100$

Processing language

Problems

- OOV: what does “wher” mean?
- word ordering

Problems

- OOV: what does “wher” mean?
- word ordering

FastText (2016, Facebook)

where = <wh, whe, her, ere, re> and <where>

wher = <wh, er>

Problems

- OOV: what does “wher” mean?
- word ordering

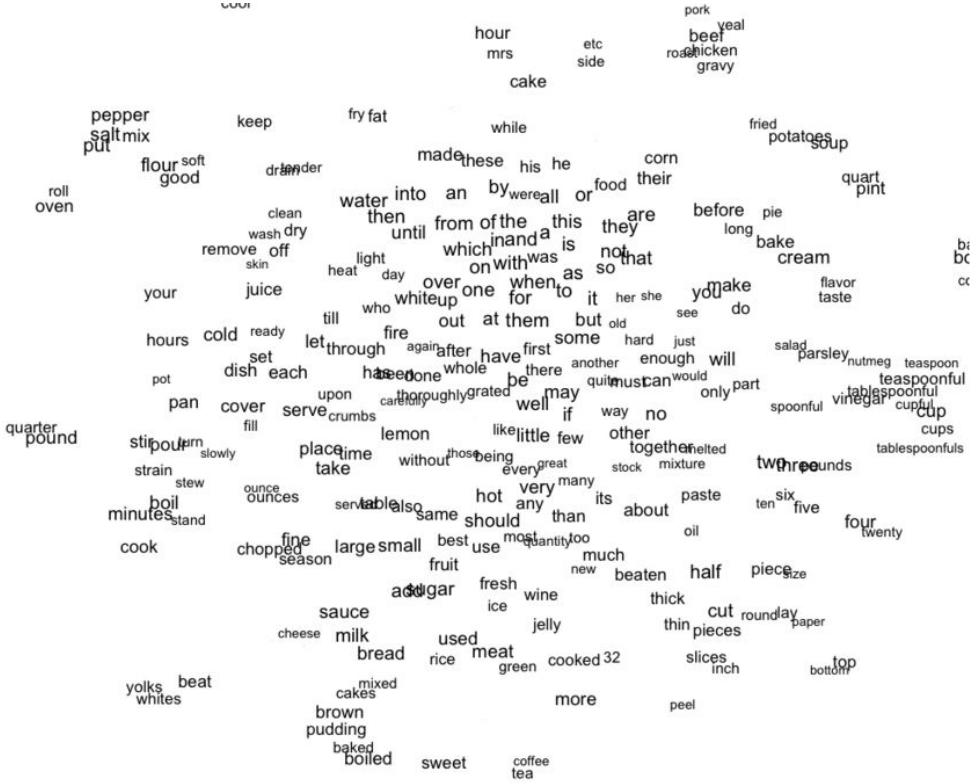
FastText (2016, Facebook)

where = <wh, whe, her, ere, re> and <where>

wher = <wh, er>

- ambiguity (context-dependency): what does “plant” mean?
- Los Angeles, Barack Obama, Doctor House

What now?


pork
hour
mrs
etc
side
beef
veal
roast
chicken
gravy
cake
pepper
keep
fry
fat
while
fried
salt
mix
put
potatoes
soup
flour
soft
made
these
his
he
corn
good
drain
colander
quart
pint
roll
oven
water
into
an
by
were
all
or
food
their
clean
then
until
from
of
the
this
are
they
before
pie
long
wash
dry
remove
off
skin
light
heat
day
which
and
is
not
that
bake
cream
on
with
was
as
so
over
one
when
for
to
it
her
she
you
make
flavor
taste
your
juice
who
white
up
see
do
till
out
at
them
but
old
hours
cold
ready
let
through
fire
again
after
have
first
some
hard
just
salad
parsley
nutmeg
teaspoon
set
dish
each
done
whole
there
another
enough
will
pot
be
quite
must
can
would
only
part
teaspoonful
tablespoonful
pan
cover
upon
serve
crumbs
carefully
thoroughly
grated
may
well
if
way
no
spoonful
vinegar
cupful
cup
quarter
pound
fill
lemon
like
little
few
other
cups
stir
pour
turn
slowly
place
lime
without
those
being
every
great
together
melted
tablespoonfuls
strain
take
stock
mixture
two
three
pounds
stew
ounce
ounces
hot
very
many
its
paste
ten
six
five
boil
minutes
stand
also
same
should
any
than
about
four
twenty
cook
chopped
fine
large
small
best
use
most
quantity
too
oil
season
fruit
much
new
beaten
half
piece
size
add
sugar
fresh
wine
thick
ice
sauce
jelly
cut
round
thin
pieces
paper
cheese
milk
used
bread
meat
rice
green
cooked
32
slices
inch
top
bottom
yolks
whites
beat
mixed
cakes
more
peel
brown
pudding
baked
boiled
sweet
coffee
tea

Operations with embeddings

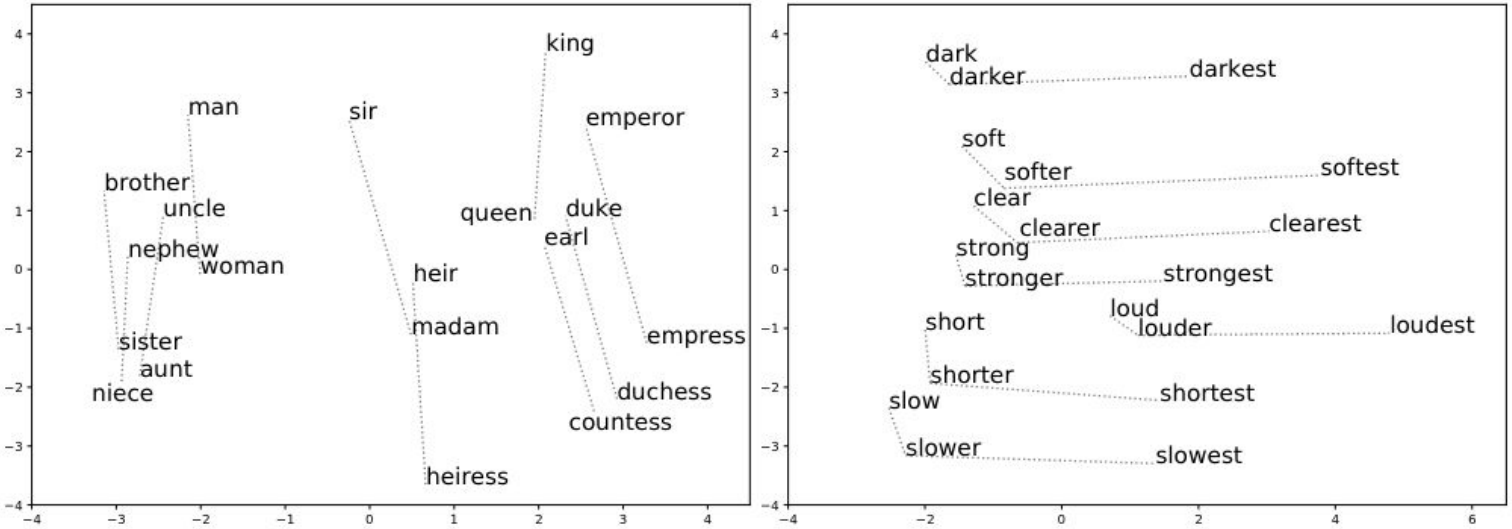
$$p = (p_1, p_2, \ldots, p_n)$$

$$q = (q_1, q_2, \ldots, q_n)$$

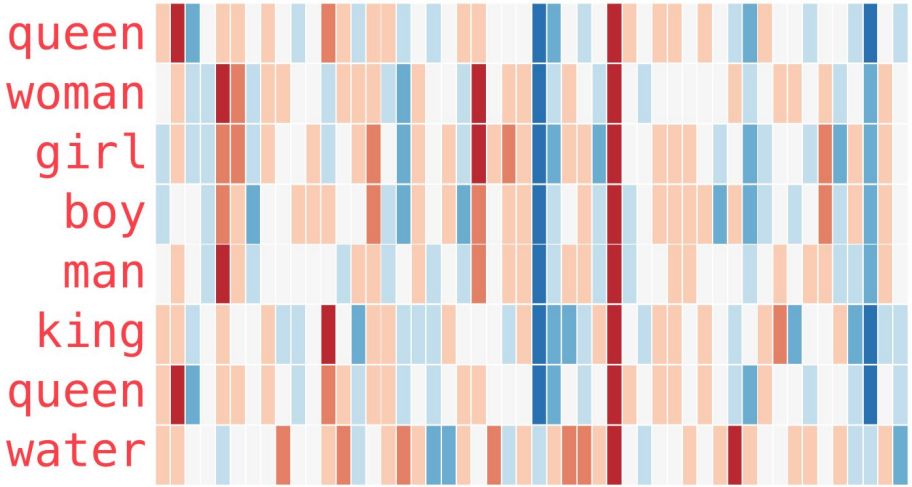
Euclidean distance $$d(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Cosine distance $$cos(p, q) = 1 - \frac{P \times Q}{\| P \|_2 \| Q \|_2} = 1 - \frac{\sum_{i=1}^{n} p_i \times q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \times \sqrt{\sum_{i=1}^{n} q_i^2}}$$

Operations with embeddings



Operations with embeddings

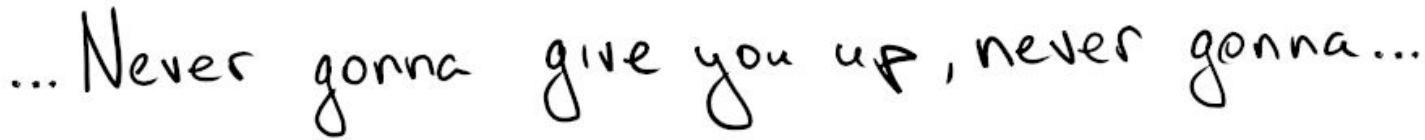


king – queen ≈ man – woman

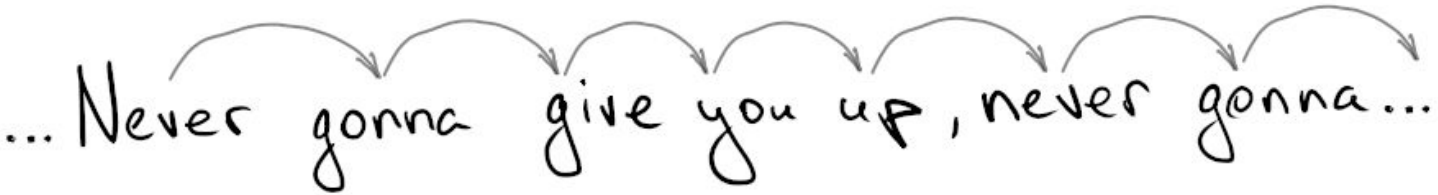
man – boy ≈? – girl
japan – sushi ≈? – pasta
euro – france ≈? – russia

Centroid(january, february, march) ≈?
Centroid(moscow, siberia, caucasus ≈?
Centroid(sunday, monday, tuesday) ≈?

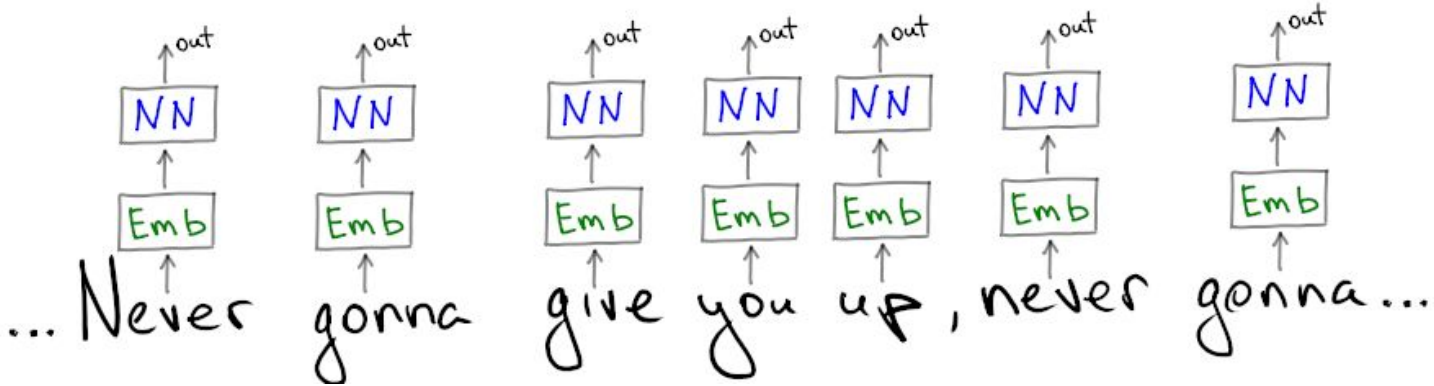
How to process sequences?



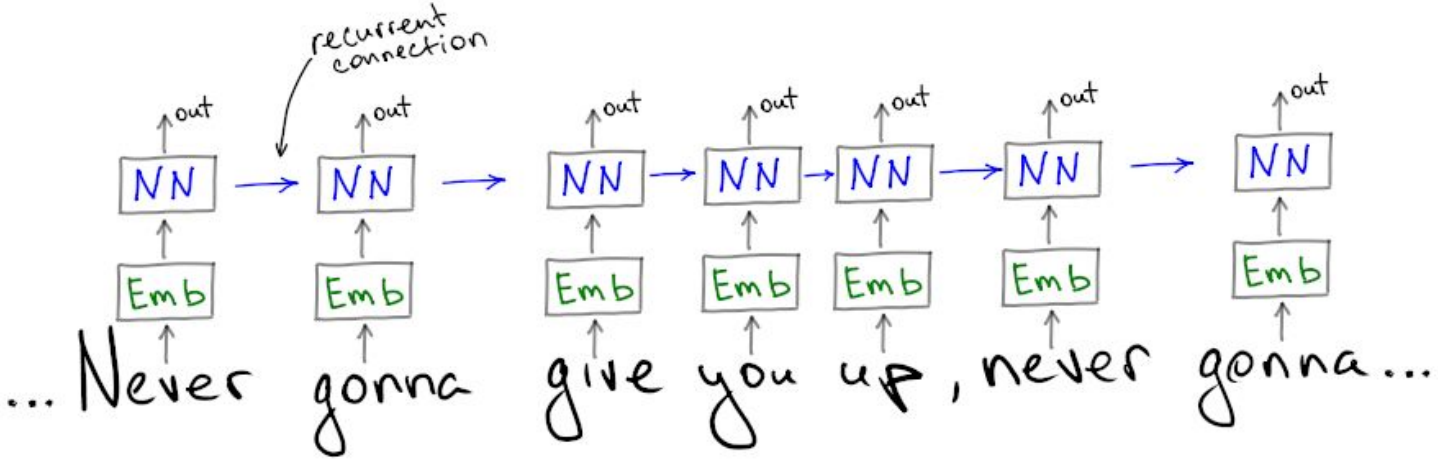
How to process sequences?



How to process sequences?



How to process sequences?



Next: RNN, LSTM

Artificial Intelligence Research Institute

airi.net

- airi_research_institute
- AIRI Institute
- AIRI Institute
- AIRI_inst
- artificial-intelligence-research-institute

Telegram

Github of NLP Course

Feedback

AIRI × Университет Сириус