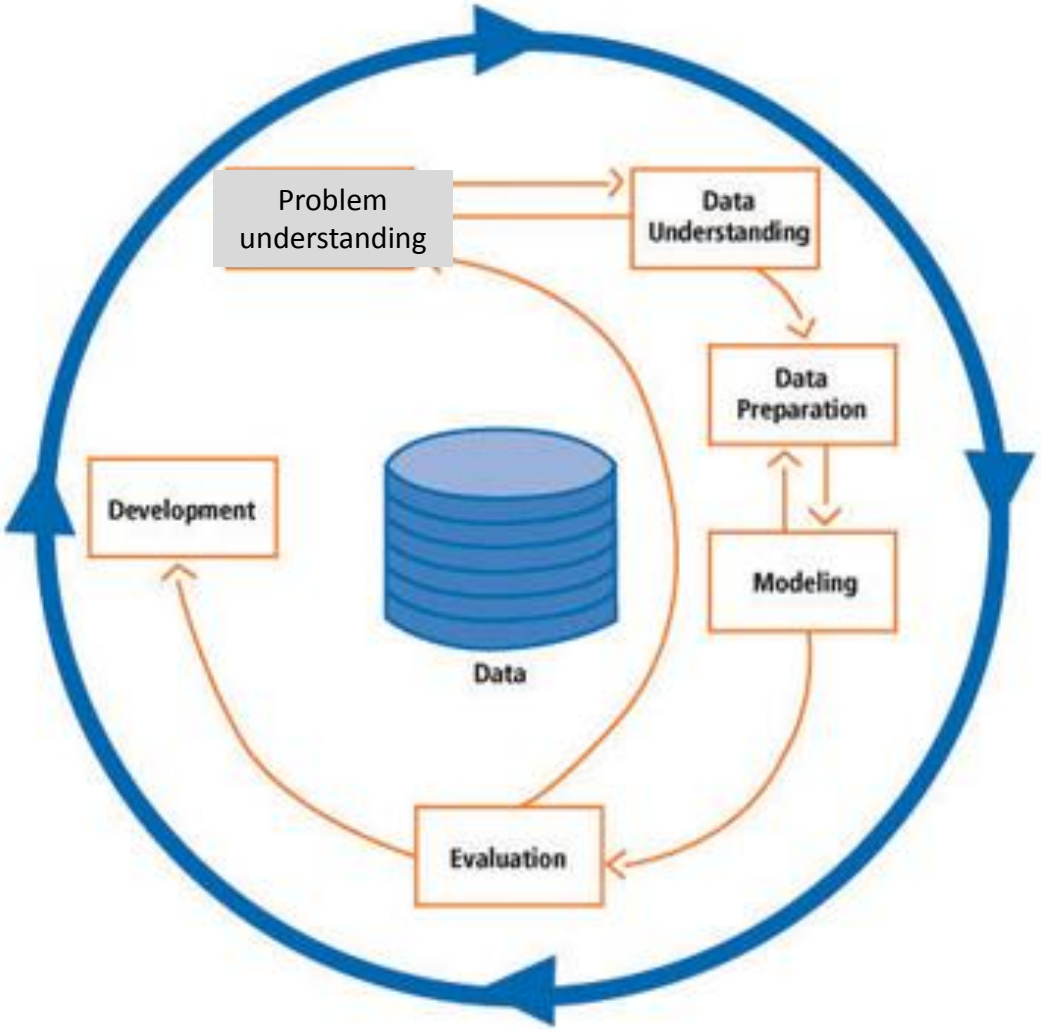
Strojno učenje

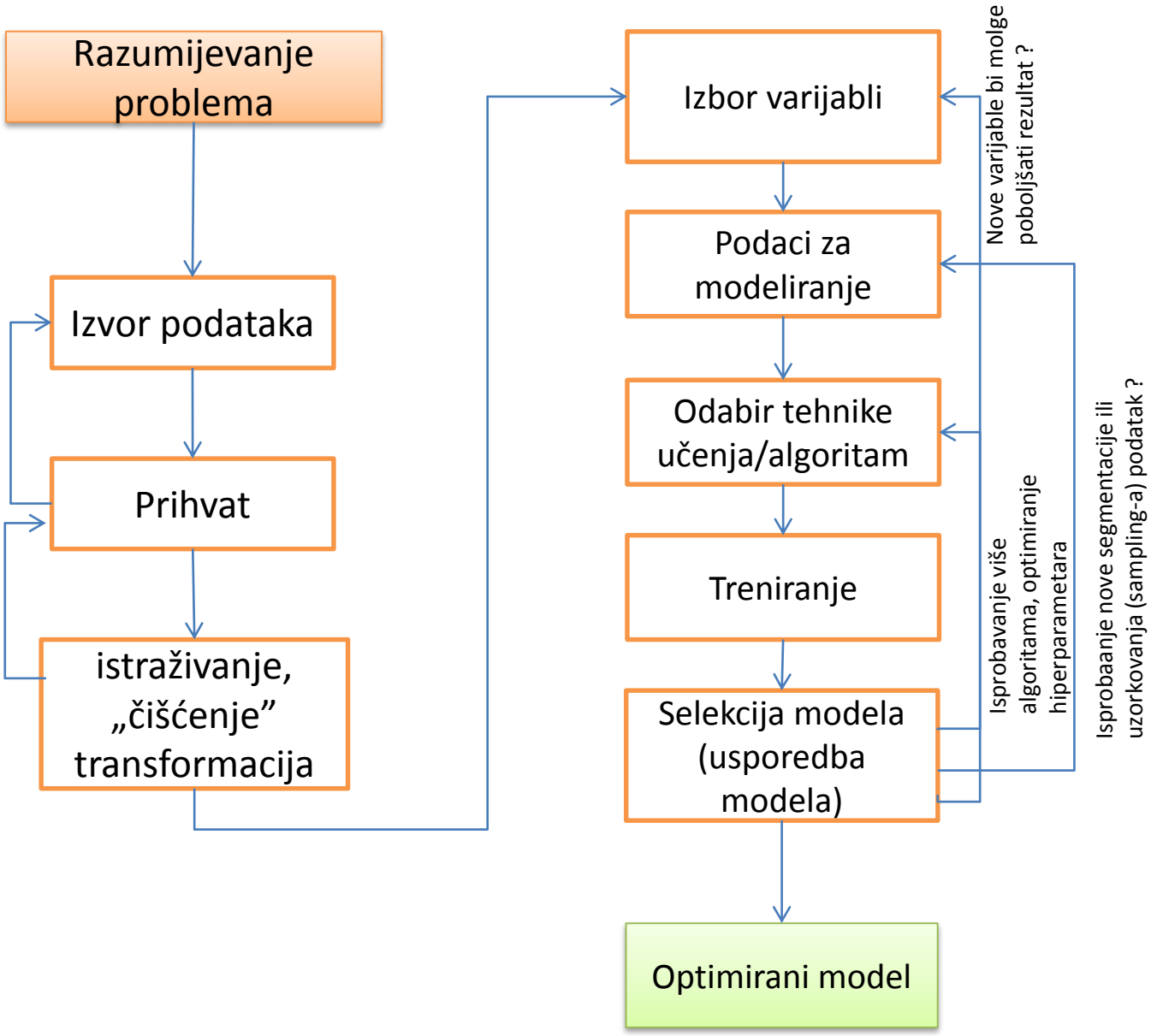
pristup problemu, praktični savjeti

Tomislav Šmuc

DM proces


Problem understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Development
Data

Prediktivno modeliranje



Koraci u rješavanju problema

Razumijevanje problema

Cilj?

- **Optimirati neki proces** odlučivanja: smanjiti trošak/štetu; povećati profit
  => uspjeh market. kampanje; online WEB-shop profit
  (prediktivni zadatak – koja je stvarna metrika uspješnosti modela)

- **Razumjeti i objasniti podatke** => koje su karakteristike ljudi oboljelih od dijabetesa – koje ih razlikuju od kontrolne skupine ?
  (deskriptivni zadatak – omogućiti razumijevanje i interpretaciju ekspertima)

- **Automatizirati neki tehnološki proces** => odobrenje transakcije; spam filtering; prepoznavanje lica/uzoraka; prevođenje
  (prediktivni zadatak – optimiranje nekog ishoda/spec.mjere;
  + zahtjevi na izvedbu (odziv u vremenu, memorijski zahtjevi)

Koraci u rješavanju problema

Razumijevanje podataka

- Značenje varijabli?
- Koje su najvažnije za naš problem? Kakve su distribucije vrijednosti ?
- Jesu li to sve varijable – trebamo li ili možemo skupiti neke druge podatke?
- Jesu li atributi/varijable korektnog tipa (kategorijske, numeričke)?
- Možemo li konstruirati nove (informativnije varijable) iz postojećih?

Koraci u rješavanju problema

Priprema podataka

Identificiranje i dohvat podataka

- Sve što je dostupno?
- U razumnom stanju (čisto)
- Podaci se smiju koristiti (komercijalna zaštita, zaštita privatnosti)
- Prikupljanje osobnih podataka (GDPR)?

- Podaci (varijable) bi trebale biti dostatni za rješavanje problema ?

  Primjer: Ako je problem identificiranje određenih klijenata/kupaca

  osnovno => podaci moraju sadržavati informacije na nivou individualnih kupaca!

- Kompletnost

  $\Rightarrow$ Problem nedostajućih vrijednosti

Koraci u rješavanju problema

Istraživanje, čišćenje, transformacija podataka

- Nedostajući podaci (MV): koliko (broj atributa s MV; % MV)
  - Metode nadomještanja MV
  - Postoje li pravilnosti u pojavnosti MV (je li njihova pojava slučajna ili pravilna?)
- Da li su vrijednosti po varijablama u razumnim granicama ?
- Da li su distribucije individualnih varijabli razumne(objašnjive)?
- Outliers – kako ih otkriti (individualne distribucije); krive vrijednosti; zamijenjene vrijednosti

Pretvaranje podataka u pravi format/ pravu reprezentaciju

- Većina algoritama radi na jednostavnim tablicama; 1 primjer- u jednom redu

  Primjer 1: relacijska baza => jednostavna tablica ?

  Primjer 2: tekstualni podaci => table (word frequency, TFIDF, word embeddings)

  Primjer 3: slike => (paketi za izradu varijabli koje opisuju sliku; podešavanje rezolucije)

  Primjer 4: vremenske serije => tablica (time windows reprezentacija)

- Vremenski podaci – određivanje granularnosti podataka;
  yy/mm/ww/dd; => agregacija vrijednosti po drugim varijablama

Koraci u rješavanju problema

Izbor reprezentacije podataka

- **Transformiranje varijabli**: procedura
  - Primjer 1: logaritamska transformacija (numeričke varijable)
  - Primjer 2: Regresijski problem se može transformirati u klasifikacijski (ako je to razumno)
  - Primjer 3: PCA/NMF (e.g. Redukcija dimenzionalnosti)

- **Konstrukcija novih varijabli**:
  - Na bazi postojećih;
  - kombinacija varijabli

    Primjeri:
    - tezina&visina => Body Mass Index
    - Export & Import => (Exp-Imp) ili Exp/Imp
  - Duboke mreže,CNN – automatsko kreiranje novih varijabli !

- **Selekcija varijabli**: samo relevantne varijable
  - korelacija sa sciljnom varijablom
  - tehnike optimiranja podskupa varijabli
  - Kauzalne varijable => kauzalni model podataka

Koraci u rješavanju problema

Modeliranje

Podaci za proces modeliranja => podaci za treniranje modela

- Distribucija podataka (npr. Brojnost Klasa?)
  - Ne-balansirani podaci:

    Primjer 1: oversampling /undersampling klasa

    Primjer 2: - cost tables (penaliziranje FP ili FN);

    - utežnjavanje primjera određene klase

Odabir validacijske tehnike

Training &Validation &Test

X-validacija

bootstrap uzorkovanje

Odabir tehnike/algoritma za modeliranje

- Različiti algoritmi koji najbolje odgovaraju za problem:
  - Prediktivni zadaci: Klasifikacija; regresija; rangiranje
  - Deskriptivni zadaci:
    - clustering (distance based/density based?)
  - otkrivanje uzoraka : Association rules, Clustering+Classification
  - Strukturiranje podatak: Hijerarhijski Clustering

Koraci u rješavanju problema

Evaluacija i odabir modela

- Procjena greške – resampling metode
  - Train and validation (model selection & hyperparam opt) + test evaluacija
  - X-validacija
  - Leave-one-out cross-validation
  - bootstrap (out-of-bag) estimation
  - Usporedba između modela

Mjere greške: klasifikacija

Bazirane na matrici konfuzije

Točnost, Osjetljivost(Recall), Preciznost, Specifičnost....

MCC – Mathew Correlation Coefficeint (za ne-balansirane podatke)

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

AUROC

AUPRC

Koraci u rješavanju problema

Mjere greške: regresija

RMSE

MAE,RMAE – Mean Absolute Error, Relative MAE

Bitno za performans: distribucija grešaka!

Problem reprezentacije za posebne strukture podataka

Posebno strukurirani i nestrukturirani podaci
Vremenske serije (sekvence)

- Forecasting – predviđanje novih vrijednosti u vremenu

- Standardni algoritmi za forecasting (R)
  - ARMA, ARIMA, GARCH, ARCH (auto-regressive modeling)

  Standardni regresijski algoritmi
  - transformiranje vremenske serije u standardnu tablicu za učenje
    => „sliding window transform”

Transformiranje vremenske serije u tablicu za ML

Date / time	Series Y
t_0	V_0
$t_0+\Delta t$	V_1
$t_0+2\Delta t$	V_2
$t_0+3\Delta t$	V_3
$t_0+4\Delta t$	V_4
$t_0+5\Delta t$	V_5
$t_0+6\Delta t$	V_6
$t_0+7\Delta t$	V_7
$t_0+8\Delta t$	V_8
…	…
$t_0+n\Delta t$	V_n

Sliding window
(windowing transform)

Sw (3,1)
(lag=3,horizon=1)

ID	x1	x2	x3	y
EX_1	V_0	V_1	V_2	V_3
EX_2	V_1	V_2	V_3	V_4
EX_3	V_2	V_3	V_4	V_5
…	…	…	…	…
EX_{n-2}	V_{n-3}	V_{n-2}	V_{n-1}	V_n

Transformiranje više vremenskih serija u tablicu za ML

Sliding window
(windowing transform)

Date / time	Series Y1	Series Y2	..	Series YN
t_0	$Y1_0$	$Y2_0$	..	YN_0
$t_0+\Delta t$	$Y1_1$	$Y2_1$	..	YN_1
$t_0+2\Delta t$	$Y1_2$	$Y2_2$	..	YN_2
$t_0+3\Delta t$	$Y1_3$	$Y2_3$	..	YN_3
$t_0+4\Delta t$	$Y1_4$	$Y2_4$	..	YN_4
$t_0+5\Delta t$	$Y1_5$	$Y2_5$	..	YN_5
$t_0+6\Delta t$	$Y1_6$	$Y2_6$	..	YN_6
$t_0+7\Delta t$	$Y1_7$	$Y2_7$	..	YN_7
$t_0+8\Delta t$	$Y1_8$	$Y2_8$	..	YN_8
…	…	…	..	…
$t_0+n\Delta t$	$Y1_n$	$Y2_n$	..	YN_n

Sw (3,1)

ID	Y1-2	Y1-1	Y2-2	Y2-1	..	YN-2	YN-1	Y1-0
EX_1	$Y1_0$	$Y1_1$	$Y2_0$	$Y2_1$	..			$Y1_2$
EX_2	$Y1_1$	$Y1_2$	$Y2_1$	$Y2_2$	..			$Y1_3$
EX_3	$Y1_2$	$Y1_3$	$Y2_2$	$Y2_3$	..			$Y1_4$
…	…	…	…	…	..			…
EX_{n-2}	$Y1_{n-3}$	$Y1_{n-2}$	$Y2_{n-1}$	$Y2_{n-1}$	..			$Y1_{n-1}$

Tekst (nestruktirirani podaci)

Platforme – paketi za TM

- WEKA, Python, R paket
- tekst u word-document frekvencije

- Primjer je dokument (tekst)
- Atribut/varijabla - riječ

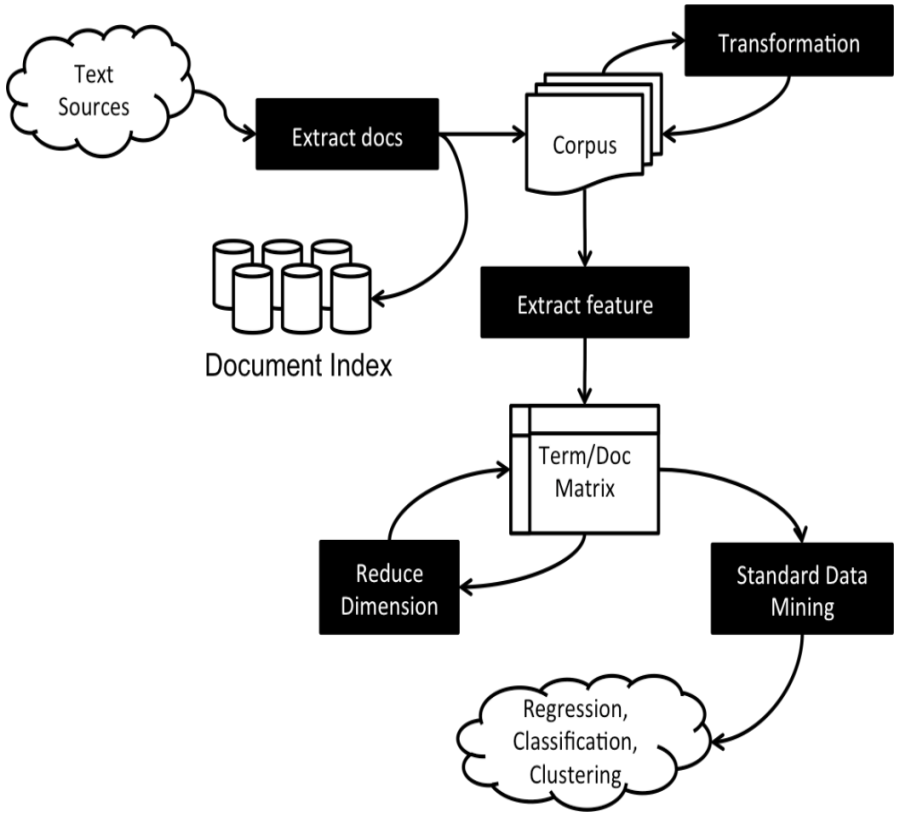
Tekst =>specifični problemi:

Linguistički problemi, prevođenje

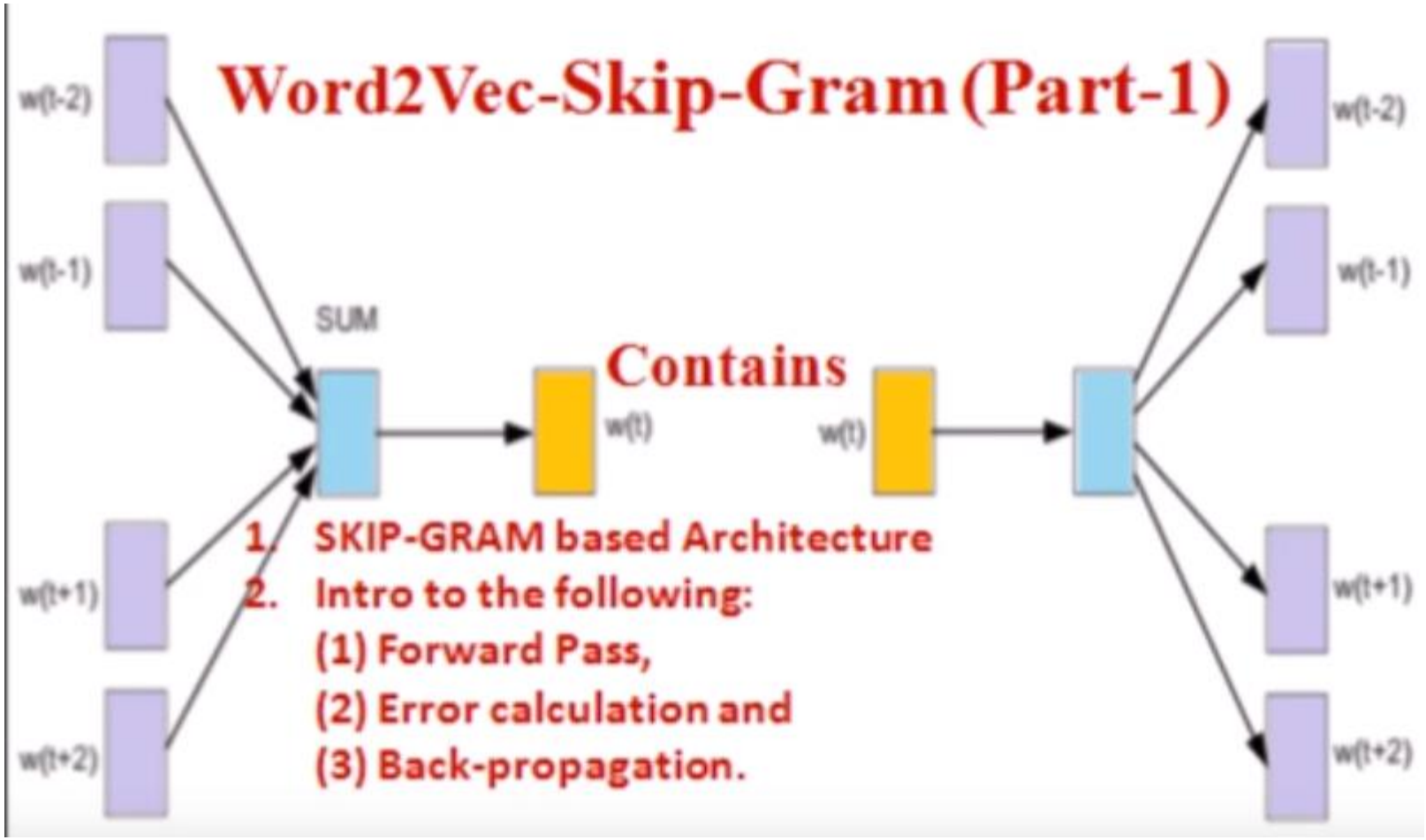
⇒ **Natural Language Processing (NLP) (**GATE**)**

⇒ **Nove tehnike – word/document embeddings (word2vec, glove)**

⇒ **Tzv. distribuirane reprezentacije riječi**

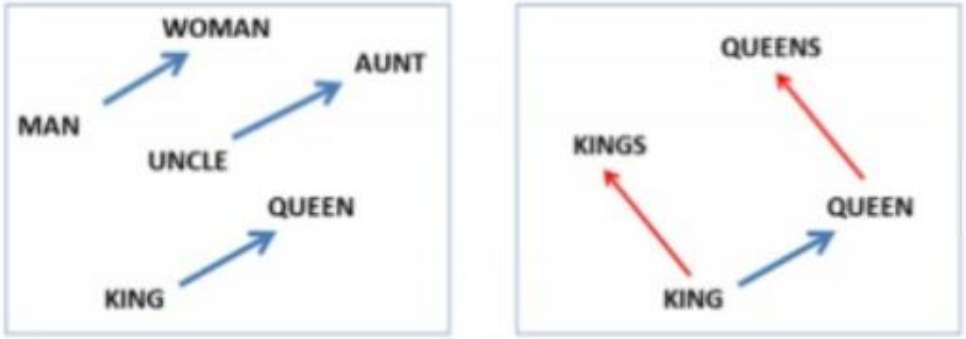


Distribuirana reprezentacija teksta
=> riječi su vektori dimenzije 50, 100...300



Distribuirana reprezentacija teksta
=> riječi su vektori dimenzije 50, 100...300

- Distributed representations for words
  - Similar words are projected into similar vectors.
  - Relationship between words can be expressed as a simple vector calculation.



[T.Mikolov et al. NIPS 2013]

- Analogy
  - v("woman") – v("man") + v("king") = v("queen")

Sudjelovanje na challenge-ima

- Data understanding
  - EDA – plot the distributions
  - Missing values
  - Outliers => consider removing before model development
  - Feature importances

- Feature engineering - or - deep learning – or- both
  - images => CNN
  - tekstovi => use distributed representations/transfer&adapt
  - time series => RNN, LSTM...
  - some features are sparse – but are they informative if combined?
  - how to combine them ?
  - PCA, NMF – transform/construct more dense representation

Sudjelovanje na challenge-u

- Model development and evaluation
  - Try simple algorithms first
  - Try different algorithms with different dataset instances (variable sets)
    - Use AUTOML to find best choice
  - Tricks: use test data in model development
    - Consider constructing validation set of the similar characteristics (distributions) => helps in diagnostics
    - Cluster all data (training and test): develop cluster models (semi-supervised learning)
  - Keep all models (reasonable models) and results => make ensembles using weighted voting schemes or some meta-learning scheme (Stacked Gen.)
  - Experience based suggestions:
    - SVM (linear) => texts
    - CNN => images
    - Mixed data => ensembles (XGBoost, Rforest)

Dodatni materijali

- Andrew Ng: Advice for applying ML (convergence, error diagnostics)
- Older competitions & reports/papers (KDD Cup)
- Kaggle competitions & solutions