kumparan

PYCON 2017 – Surabaya, Indonesia

Word2vec :
Ketika Mesin Mengerti Bahasa Manusia

Politeknik Elektronika Negeri Surabaya
Surabaya, 9 Oktober 2017

kumparan

Perkenalan

- Nama : Afif Akbar Iskandar
- Email : afifai@sci.ui.ac.id
- Bidang : Machine Learning, Deep Learning, Computer Vision
- Institusi : Kumparan (PT. Dynamo Media Network)
- Role : Data Scientist
- Blog : http://afifai.com

kumparan

Word2vec, apa itu ?

- Word Embeddings
- Pemetaan kata terhadap ruang semantik/sintatikal
- Berbasis Neural Network

Namun :

- Bukan Deep Learning

kumparan

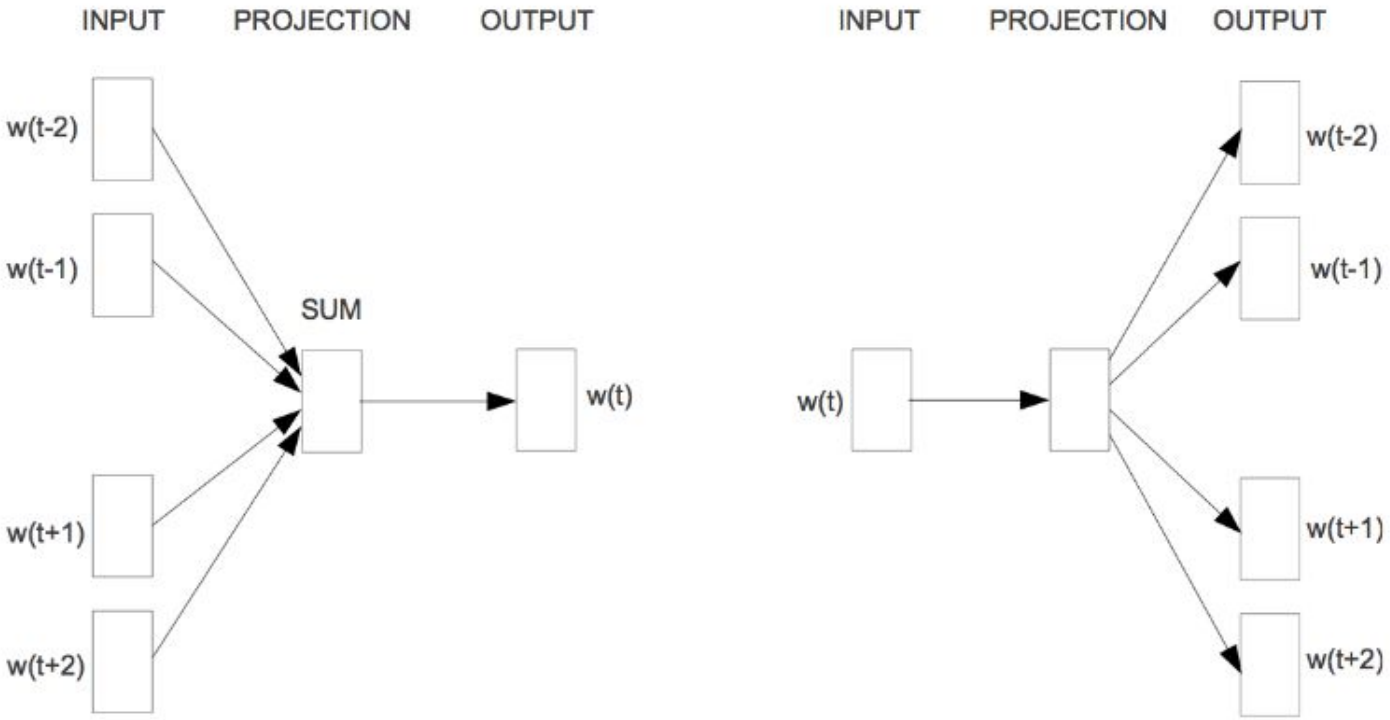
Arsitektur

- Skip-gram
- Continuous Bag of Words (CBOW)

Keduanya tidak memiliki *hidden layer*, melainkan *projection layer*

kumparan

Arsitektur



kumparan

Kelebihan

- Scales
  - Dilatih dengan menggunakan milyaran korpora dalam waktu yang cukup singkat
  - Dapat menggunakan *online learning*
- Model dapat digunakan dalam berbagai task
  - Klasifikasi Artikel
  - Pembangkit Kalimat
- Mudah dibuat menggunakan Python
  - Gensim

kumparan

Apa yang Bisa Dilakukan?

Kata-kata direpresentasikan kedalam Vektor, sehingga dapat dioperasikan dengan operasi Vektor.

Dengan korpus yang bagus (e.g.:Wikipedia)

‘King’ + ‘Woman’ - ‘Man’ = ‘Queen’

kumparan

Apa yang Bisa Dilakukan?

```
In [17]: model.most_similar(positive=[ 'presiden' , 'wanita' ], negative=[ 'pria' ])
Out[17]:
[('kepresidenan', 0.5164607167243958),
 ('presidennya', 0.5102983713150024),
 ('wapres', 0.443649023771286),
 ('soekarnoputri', 0.434302806854248O5),
 ('menlu', 0.4306909441947937),
 ('kanselir', 0.41026079654693604),
 ('macapagal', 0.403542280197143S5),
 ('megawati', 0.39232367277145386),
 ('mbeki', 0.3865049183368683),
 ('disumpah', 0.3826873302459717)]
```

kumparan

Apa yang Bisa Dilakukan?

```
In [19]: model.most_similar('surabaya')
Out[19]:
[('malang', 0.6218435168266296),
 ('semarang', 0.5621165037155151),
 ('sidoarjo', 0.5270854234695435),
 ('jogjakarta', 0.5220928192138672),
 ('madiun', 0.5171178579330444),
 ('mojokerto', 0.5162099003791809),
 ('jatim', 0.5134848952293396),
 ('gresik', 0.5091941356658936),
 ('jember', 0.495974510908126 83),
 ('kenjeran', 0.4852325916290283)]
```

kumparan

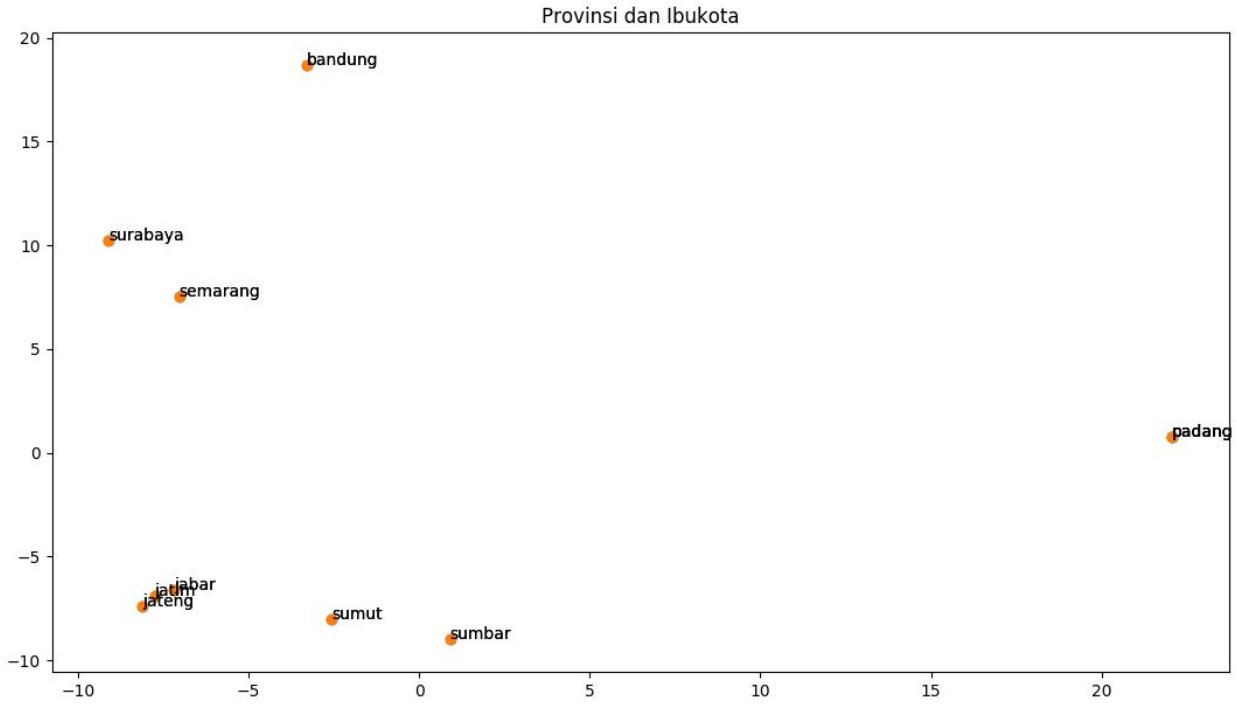
Apa yang Bisa Dilakukan?

```
In [22]: model.doesnt_match('jokowi prabowo jk pisang'.split())
Out[22]: 'pisang'

In [23]: model.doesnt_match('jambu mangga novanto pisang'.split())
Out[23]: 'novanto'
```

kumparan

Visualisasi Vektor



kumparan

Evaluasi Word2Vec

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

kumparan

Melatih Word2Vec dengan Gensim

```
import multiprocessing
from gensim.corpora.wikicorpus import WikiCorpus
from gensim.models.word2vec import Word2Vec

wiki = WikiCorpus('idwiki-latest-pages-articles.xml.bz2',
                  lemmatize=False, dictionary={})
sentences = list(wiki.get_texts())
params = {'size': 500, 'window': 10, 'min_count': 10,
          'workers': max(1, multiprocessing.cpu_count() - 1), 'sample': 1E-3,}
word2vec = Word2Vec(sentences, **params)
word2vec.save('idwiki500')
```

kumparan

Aplikasi yang dapat dibangun

- *Synonym Detection*
- *Selecting Out-of-list Words*
- *Sentence Completion*
- *Article Classification*
- *Machine Translation*
- *Auto Image Captioning*
- *Chatbot*
- *etc.*

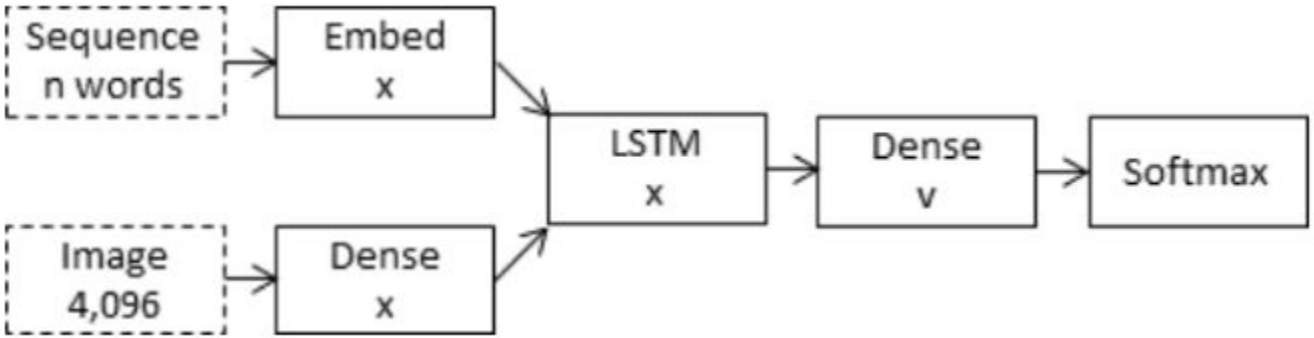
kumparan

Kekurangan Word2vec

- Ambiguitas, misalkan kata : bunga, memiliki 2 arti
- *Ignore* kata yang tidak ditemukan di data Train
- Penggunaan parameter cenderung *gambling* (Berdasarkan *trial and error*)

kumparan

Studi Kasus : Image Captioning



Studi Kasus : Image Captioning

APLIKASI PEMBANGKIT CAPTION OTOMATIS

seorang pengendara motor yang mengenakan seragam biru dan putih sedang mengendarai sepeda motor .

kumparan

Referensi

- https://radimrehurek.com/gensim/models/word2vec.html
- http://machinelearningmastery.com

kumparan

Thank You