# word2vec Demo

Efficient Estimation of Word Representations in Vector Space (Mikolov et al, 2013).

Andres Calderon and Hinna Shabir

University of California, Riverside

November 18, 2015

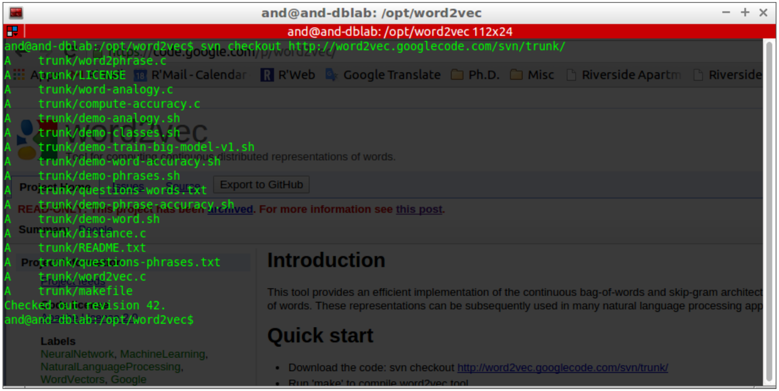
## Agenda

## Agenda

## word2vec source code

- `https://code.google.com/p/word2vec/`.
- Provides an efficient implementation of the continuous bag-of-words and skip-gram.
- Clean and well documented code in C.

## word2vec source code



and@and-dblab: /opt/word2vec

```
and@and-dblab:/opt/word2vec$ svn checkout http://word2vec.googlecode.com/svn/trunk/
A    trunk/word2phrase.c
A    trunk/LICENSE
A    trunk/word-analogy.c
A    trunk/compute-accuracy.c
A    trunk/demo-analogy.sh
A    trunk/demo-classes.sh
A    trunk/demo-train-big-model-v1.sh
A    trunk/demo-word-accuracy.sh
A    trunk/demo-phrases.sh
A    trunk/questions-words.txt
A    trunk/demo-phrase-accuracy.sh
A    trunk/demo-word.sh
A    trunk/distance.c
A    trunk/README.txt
A    trunk/questions-phrases.txt
A    trunk/word2vec.c
A    trunk/makefile
Checked out revision 42.
and@and-dblab:/opt/word2vec$
```

Introduction

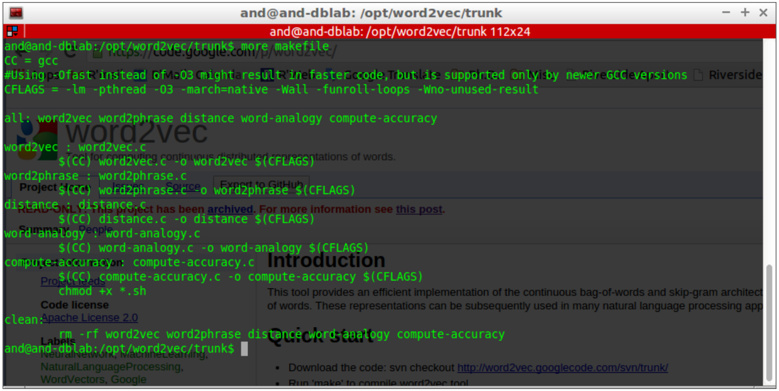
This tool provides an efficient implementation of the continuous bag-of-words and skip-gram architect of words. These representations can be subsequently used in many natural language processing app

Quick start

- Download the code: svn checkout http://word2vec.googlecode.com/svn/trunk/
- Run 'make' to compile word2vec tool

and@and-dblab: /opt/word2vec 112x24
R'Mail - Calendar
R'Web
Google Translate
Ph.D.
Misc
Riverside Apartm
Riverside
word2vec
Tool for computing continuous distributed representations of words.
Export to GitHub
READ-ONLY: This project has been archived. For more information see this post.
Labels
NeuralNetwork, MachineLearning, NaturalLanguageProcessing, WordVectors, Google

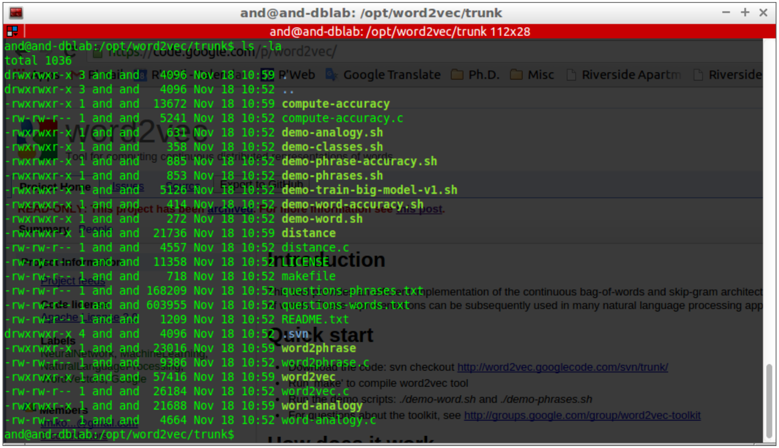
## word2vec source code



```
and@and-dblab:/opt/word2vec/trunk$ more makefile
CC = gcc
#Using -Ofast instead of -O3 might result in faster code, but is supported only by newer GCC versions
CFLAGS = -lm -pthread -O3 -march=native -Wall -funroll-loops -Wno-unused-result

all: word2vec word2phrase distance word-analogy compute-accuracy

word2vec : word2vec.c
	$(CC) word2vec.c -o word2vec $(CFLAGS)
word2phrase : word2phrase.c
	$(CC) word2phrase.c -o word2phrase $(CFLAGS)
distance : distance.c
	$(CC) distance.c -o distance $(CFLAGS)
word-analogy : word-analogy.c
	$(CC) word-analogy.c -o word-analogy $(CFLAGS)
compute-accuracy : compute-accuracy.c
	$(CC) compute-accuracy.c -o compute-accuracy $(CFLAGS)
	chmod +x *.sh

clean:
	rm -rf word2vec word2phrase distance word-analogy compute-accuracy
and@and-dblab:/opt/word2vec/trunk$
```

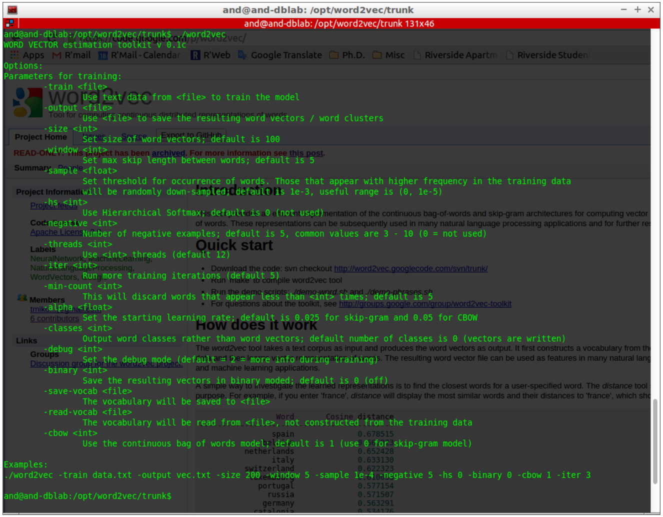
and@and-dblab: /opt/word2vec/trunk
and@and-dblab: /opt/word2vec/trunk 112x24

## word2vec source code


and@and-dblab: /opt/word2vec/trunk
and@and-dblab: /opt/word2vec/trunk 112x28

```
and@and-dblab:/opt/word2vec/trunk$ ls -la
total 1036
drwxrwxr-x 3 and and   4096 Nov 18 10:59 .
drwxrwxr-x 3 and and   4096 Nov 18 10:52 ..
-rwxrwxr-x 1 and and  13672 Nov 18 10:59 compute-accuracy
-rw-rw-r-- 1 and and   5241 Nov 18 10:52 compute-accuracy.c
-rwxrwxr-x 1 and and    631 Nov 18 10:52 demo-analogy.sh
-rwxrwxr-x 1 and and    358 Nov 18 10:52 demo-classes.sh
-rwxrwxr-x 1 and and    885 Nov 18 10:52 demo-phrase-accuracy.sh
-rwxrwxr-x 1 and and    853 Nov 18 10:52 demo-phrases.sh
-rwxrwxr-x 1 and and   5126 Nov 18 10:52 demo-train-big-model-v1.sh
-rwxrwxr-x 1 and and    414 Nov 18 10:52 demo-word-accuracy.sh
-rwxrwxr-x 1 and and    272 Nov 18 10:52 demo-word.sh
-rwxrwxr-x 1 and and  21736 Nov 18 10:59 distance
-rw-rw-r-- 1 and and   4557 Nov 18 10:52 distance.c
-rw-rw-r-- 1 and and  11358 Nov 18 10:52 LICENSE
-rw-rw-r-- 1 and and    718 Nov 18 10:52 makefile
-rw-rw-r-- 1 and and 168209 Nov 18 10:52 questions-phrases.txt
-rw-rw-r-- 1 and and 603955 Nov 18 10:52 questions-words.txt
-rw-rw-r-- 1 and and   1209 Nov 18 10:52 README.txt
drwxrwxr-x 4 and and   4096 Nov 18 10:52 .svn
-rwxrwxr-x 1 and and  23016 Nov 18 10:59 word2phrase
-rw-rw-r-- 1 and and   9386 Nov 18 10:52 word2phrase.c
-rwxrwxr-x 1 and and  57416 Nov 18 10:59 word2vec
-rw-rw-r-- 1 and and  26184 Nov 18 10:52 word2vec.c
-rwxrwxr-x 1 and and  21688 Nov 18 10:59 word-analogy
-rw-rw-r-- 1 and and   4664 Nov 18 10:52 word-analogy.c
and@and-dblab:/opt/word2vec/trunk$
```

## word2vec source code

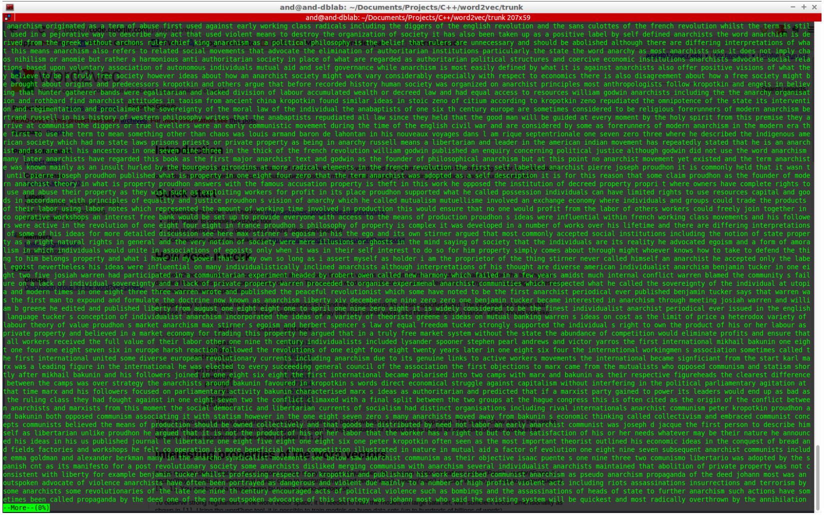

and@and-dblab: /opt/word2vec/trunk
and@and-dblab: /opt/word2vec/trunk 131x46

```
and@and-dblab:/opt/word2vec/trunk$ ./word2vec
WORD VECTOR estimation toolkit v 0.1c

Options:
Parameters for training:
        -train <file>
                Use text data from <file> to train the model
        -output <file>
                Use <file> to save the resulting word vectors / word clusters
        -size <int>
                Set size of word vectors; default is 100
        -window <int>
                Set max skip length between words; default is 5
        -sample <float>
                Set threshold for occurrence of words. Those that appear with higher frequency in the training data
                will be randomly down-sampled; default is 1e-3, useful range is (0, 1e-5)
        -hs <int>
                Use Hierarchical Softmax; default is 0 (not used)
        -negative <int>
                Number of negative examples; default is 5, common values are 3 - 10 (0 = not used)
        -threads <int>
                Use <int> threads (default 12)
        -iter <int>
                Run more training iterations (default 5)
        -min-count <int>
                This will discard words that appear less than <int> times; default is 5
        -alpha <float>
                Set the starting learning rate; default is 0.025 for skip-gram and 0.05 for CBOW
        -classes <int>
                Output word classes rather than word vectors; default number of classes is 0 (vectors are written)
        -debug <int>
                Set the debug mode (default = 2 = more info during training)
        -binary <int>
                Save the resulting vectors in binary moded; default is 0 (off)
        -save-vocab <file>
                The vocabulary will be saved to <file>
        -read-vocab <file>
                The vocabulary will be read from <file>, not constructed from the training data
        -cbow <int>
                Use the continuous bag of words model; default is 1 (use 0 for skip-gram model)

Examples:
./word2vec -train data.txt -output vec.txt -size 200 -window 5 -sample 1e-4 -negative 5 -hs 0 -binary 0 -cbow 1 -iter 3

and@and-dblab:/opt/word2vec/trunk$
```

## demo-word.sh

```
## Compile the code...
make
## Download and unzip the training file...
if [ ! -e text8 ]; then
  wget http://mattmahoney.net/dc/text8.zip -O text8.gz
  gzip -d text8.gz -f
fi
## Run the model (taking time)...
time ./word2vec -train text8 -output vectors.bin -cbow 1 -size 200 -window 8 -negative 25 -hs 0
  ↪    -sample 1e-4 -threads 20 -binary 1 -iter 15
## Query word distances...
./distance vectors.bin
```

## text8 file



```
and@and-dblab: ~/Documents/Projects/C++/word2vec/trunk
and@and-dblab: ~/Documents/Projects/C++/word2vec/trunk 207x59
 anarchism originated as a term of abuse first used against early working class radicals including the diggers of the english revolution and the sans culottes of the french revolution whilst the term is stil
l used in a pejorative way to describe any act that used violent means to destroy the organization of society it has also been taken up as a positive label by self defined anarchists the word anarchism is de
rived from the greek without archons ruler chief king anarchism as a political philosophy is the belief that rulers are unnecessary and should be abolished although there are differing interpretations of wha
t this means anarchism also refers to related social movements that advocate the elimination of authoritarian institutions particularly the state the word anarchy as most anarchists use it does not imply cha
os nihilism or anomie but rather a harmonious anti authoritarian society in place of what are regarded as authoritarian political structures and coercive economic institutions anarchists advocate social rela
tions based upon voluntary association of autonomous individuals mutual aid and self governance while anarchism is most easily defined by what it is against anarchists also offer positive visions of what the
y believe to be a truly free society however ideas about how an anarchist society might work vary considerably especially with respect to economics there is also disagreement about how a free society might b
e brought about origins and predecessors kropotkin and others argue that before recorded history human society was organized on anarchist principles most anthropologists follow kropotkin and engels in believ
ing that hunter gatherer bands were egalitarian and lacked division of labour accumulated wealth or decreed law and had equal access to resources william godwin anarchists including the the anarchy organisat
ion and rothbard find anarchist attitudes in taoism from ancient china kropotkin found similar ideas in stoic zeno of citium according to kropotkin zeno repudiated the omnipotence of the state its interventi
on and regimentation and proclaimed the sovereignty of the moral law of the individual the anabaptists of one six th century europe are sometimes considered to be religious forerunners of modern anarchism be
rtrand russell in his history of western philosophy writes that the anabaptists repudiated all law since they held that the good man will be guided at every moment by the holy spirit from this premise they a
rrive at communism the diggers or true levellers were an early communistic movement during the time of the english civil war and are considered by some as forerunners of modern anarchism in the modern era th
e first to use the term to mean something other than chaos was louis armand baron de lahontan in his nouveaux voyages dans l am rique septentrionale one seven zero three where he described the indigenous ame
rican society which had no state laws prisons priests or private property as being in anarchy russell means a libertarian and leader in the american indian movement has repeatedly stated that he is an anarch
ist and so are all his ancestors in one seven nine three in the thick of the french revolution william godwin published an enquiry concerning political justice although godwin did not use the word anarchism
many later anarchists have regarded this book as the first major anarchist text and godwin as the founder of philosophical anarchism but at this point no anarchist movement yet existed and the term anarchist
e was known mainly as an insult hurled by the bourgeois girondins at more radical elements in the french revolution the first self labelled anarchist pierre joseph proudhon it is commonly held that it wasn t
 until pierre joseph proudhon published what is property in one eight four zero that the term anarchist was adopted as a self description it is for this reason that some claim proudhon as the founder of mode
rn anarchist theory in what is property proudhon answers with the famous accusation property is theft in this work he opposed the institution of decreed property propri t where owners have complete rights to
 use and abuse their property as they wish such as exploiting workers for profit in its place proudhon supported what he called possession individuals can have limited rights to use resources capital and goo
ds in accordance with principles of equality and justice proudhon s vision of anarchy which he called mutualism mutuellisme involved an exchange economy where individuals and groups could trade the products
of their labor using labor notes which represented the amount of working time involved in production this would ensure that no one would profit from the labor of others workers could freely join together in
co operative workshops an interest free bank would be set up to provide everyone with access to the means of production proudhon s ideas were influential within french working class movements and his followe
rs were active in the revolution of one eight four eight in france proudhon s philosophy of property is complex it was developed in a number of works over his lifetime and there are differing interpretations
 of some of his ideas for more detailed discussion see here max stirner s egoism in his the ego and its own stirner argued that most commonly accepted social institutions including the notion of state proper
ty as a right natural rights in general and the very notion of society were mere illusions or ghosts in the mind saying of society that the individuals are its reality he advocated egoism and a form of amora
lism in which individuals would unite in associations of egoists only when it was in their self interest to do so for him property simply comes about through might whoever knows how to take to defend the thi
ng to him belongs property and what i have in my power that is my own so long as i assert myself as holder i am the proprietor of the thing stirner never called himself an anarchist he accepted only the labe
l egoist nevertheless his ideas were influential on many individualistically inclined anarchists although interpretations of his thought are diverse american individualist anarchism benjamin tucker in one ei
ght two five josiah warren had participated in a communitarian experiment headed by robert owen called new harmony which failed in a few years amidst much internal conflict warren blamed the community s fail
ure on a lack of individual sovereignty and a lack of private property warren proceeded to organise experimenal anarchist communities which respected what he called the sovereignty of the individual at utopi
a and modern times in one eight three three warren wrote and published the peaceful revolutionist which some have noted to be the first anarchist periodical ever published benjamin tucker says that warren wa
s the first man to expound and formulate the doctrine now known as anarchism liberty xiv december one nine zero zero one benjamin tucker became interested in anarchism through meeting josiah warren and willi
am b greene he edited and published liberty from august one eight eight one to april one nine zero eight it is widely considered to be the finest individualist anarchist periodical ever issued in the english
 language tucker s conception of individualist anarchism incorporated the ideas of a variety of theorists greene s ideas on mutual banking warren s ideas on cost as the limit of price a heterodox variety of
labour theory of value proudhon s market anarchism max stirner s egoism and herbert spencer s law of equal freedom tucker strongly supported the individual s right to own the product of his or her labour as
private property and believed in a market economy for trading this property he argued that in a truly free market system without the state the abundance of competition would eliminate profits and ensure that
 all workers received the full value of their labor other one nine th century individualists included lysander spooner stephen pearl andrews and victor yarros the first international mikhail bakunin one eigh
t one four one eight seven six in europe harsh reaction followed the revolutions of one eight four eight twenty years later in one eight six four the international workingmen s association sometimes called t
he first international united some diverse european revolutionary currents including anarchism due to its genuine links to active workers movements the international became signficiant from the start karl ma
rx was a leading figure in the international he was elected to every succeeding general council of the association the first objections to marx came from the mutualists who opposed communism and statism shor
tly after mikhail bakunin and his followers joined in one eight six eight the first international became polarised into two camps with marx and bakunin as their respective figureheads the clearest difference
 between the camps was over strategy the anarchists around bakunin favoured in kropotkin s words direct economical struggle against capitalism without interfering in the political parliamentary agitation at
that time marx and his followers focused on parliamentary activity bakunin characterised marx s ideas as authoritarian and predicted that if a marxist party gained to power its leaders would end up as bad as
 the ruling class they had fought against in one eight seven two the conflict climaxed with a final split between the two groups at the hague congress this is often cited as the origin of the conflict betwee
n anarchists and marxists from this moment the social democratic and libertarian currents of socialism had distinct organisations including rival internationals anarchist communism peter kropotkin proudhon a
nd bakunin both opposed communism associating it with statism however in the one eight seven zero s many anarchists moved away from bakunin s economic thinking called collectivism and embraced communist conc
epts communists believed the means of production should be owned collectively and that goods be distributed by need not labor an early anarchist communist was joseph d jacque the first person to describe him
self as libertarian unlike proudhon he argued that it is not the product of his or her labor that the worker has a right to but to the satisfaction of his or her needs whatever may be their nature he announc
ed his ideas in his us published journal le libertaire one eight five eight one eight six one peter kropotkin often seen as the most important theorist outlined his economic ideas in the conquest of bread an
d fields factories and workshops he felt co operation is more beneficial than competition illustrated in nature in mutual aid a factor of evolution one eight nine seven subsequent anarchist communists includ
e emma goldman and alexander berkman many in the anarcho syndicalist movements see below saw anarchist communism as their objective isaac puente s one nine three two comunismo libertario was adopted by the s
panish cnt as its manifesto for a post revolutionary society some anarchists disliked merging communism with anarchism several individualist anarchists maintained that abolition of private property was not c
onsistent with liberty for example benjamin tucker whilst professing respect for kropotkin and publishing his work described communist anarchism as pseudo anarchism propaganda of the deed johann most was an
outspoken advocate of violence anarchists have often been portrayed as dangerous and violent due mainly to a number of high profile violent acts including riots assassinations insurrections and terrorism by
some anarchists some revolutionaries of the late one nine th century encouraged acts of political violence such as bombings and the assassinations of heads of state to further anarchism such actions have som
etimes been called propaganda by the deed one of the more outspoken advocates of this strategy was johann most who said the existing system will be quickest and most radically overthrown by the annihilation
--More--(0%)
```

## demo-word.sh output

```
and@and-dblab:~/Documents/Projects/C++/word2vec/trunk$ ./demo-word.sh
make: Nothing to be done for 'all'.
--2015-11-12 18:18:10--  http://mattmahoney.net/dc/text8.zip
Resolving mattmahoney.net (mattmahoney.net)... 98.139.135.129
Connecting to mattmahoney.net (mattmahoney.net)|98.139.135.129|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 31344016 (30M) [application/zip]
Saving to: text8.gz

text8.gz
  ↪    100%[==================================================>]  29.89M  1.74MB/s   in 18s

2015-11-12 18:18:28 (1.70 MB/s) - text8.gz saved [31344016/31344016]

Starting training using file text8
Vocab size: 71291
Words in train file: 16718843
Alpha: 0.000005  Progress: 100.10%  Words/thread/sec: 113.47k
real  10m15.450s
user  36m52.552s
sys  0m4.388s
Enter word or sentence (EXIT to break):
```

## demo-00.sh

```
## Get a small file...
head -c 5000000 text8 > text8_small
## Build the model...
./word2vec -train text8_small -output vectors_small.bin -cbow 1 -size 100 -window 5 -negative 0 -hs
  ↪    25 -threads 1 -iter 4 -min-count 2 -binary 1
## Query word distances...
./distance vectors_small.bin
```

## demo-01.sh

```
## Text model saving vocabulary
./word2vec -train text8_small -output vectors_small_50.txt -cbow 1 -size 50 -window 5 -negative 0
  ↪    -hs 25 -threads 1 -iter 4 -binary 0 -save-vocab vocab.txt
## Text model with just 3 dimensions
./word2vec -train text8_small -output vectors_small_3.txt -cbow 1 -size 3 -window 5 -negative 0 -hs
  ↪    25 -threads 1 -iter 4 -binary 0
## See the results...
echo "Text model size 50..."
head -n 5 vectors_small_50.txt
echo "Vocabulary..."
head -n 5 vocab.txt
echo "Text model size 3..."
head -n 5 vectors_small_3.txt
```

## demo-word.sh revisited

- **distance** can load a pre-trained model...
- Let's try some examples...
  1. california
  2. sciences
  3. happiness
  4. man
  5. ...

## demo-classes.sh

```
## Same than before...
make
if [ ! -e text8 ]; then
  wget http://mattmahoney.net/dc/text8.zip -O text8.gz
  gzip -d text8.gz -f
fi
## Train the model with classes rather than vectors...
time ./word2vec -train text8 -output classes.txt -cbow 1 -size 200 -window 8 -negative 25 -hs 0
  ↪    -sample 1e-4 -threads 20 -iter 15 -classes 500
## Sort the result by the second column...
sort classes.txt -k 2 -n > classes.sorted.txt
echo The word classes were saved to file classes.sorted.txt
```

## Interesting properties of the word vectors

- $\overrightarrow{paris} - \overrightarrow{france} + \overrightarrow{italy} \cong \overrightarrow{rome}$
- $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{women} \cong \overrightarrow{queen}$

## demo-analogy.sh

```
## Same that before...
make
if [ ! -e text8 ]; then
  wget http://mattmahoney.net/dc/text8.zip -O text8.gz
  gzip -d text8.gz -f
fi
echo --------------------------------------------------------
echo Note that for the word analogy to perform well, the model
echo should be trained on much larger data set
echo Example input: paris france berlin
echo --------------------------------------------------------
time ./word2vec -train text8 -output vectors.bin -cbow 1 -size 200 -window 8 -negative 25 -hs 0
  ↪    -sample 1e-4 -threads 20 -binary 1 -iter 15
## Call word-analogy script...
./word-analogy vectors.bin
```

## demo-analogy.sh

- Some examples...
  1. paris france bogota ...
  2. boy girl brother ...
  3. chicago illinois memphis ...
  4. poland zloty sweden ...
  5. bad worst good ...
  6. child children mouse ...
  7. going went selling ...
  8. king man queen ...
  9. mexico mexican peru ...
  10. berlin germany riyadh[1] ...
  11. woman angel man ...

[1] **word2phrase** will address the problem...

## demo-phrases.sh

```
## Compile...
make
## Download...
if [ ! -e news.2012.en.shuffled ]; then
  wget http://www.statmt.org/wmt14/training-monolingual-news-crawl/news.2012.en.shuffled.gz
  gzip -d news.2012.en.shuffled.gz -f
fi
## Pre-process...
sed -e "s//’/g" -e "s//’/g" -e "s/’’/ /g" < news.2012.en.shuffled | tr -c "A-Za-z’_ \n" " " >
  ↪    news.2012.en.shuffled-norm0
time ./word2phrase -train news.2012.en.shuffled-norm0 -output news.2012.en.shuffled-norm0-phrase0
  ↪    -threshold 200 -debug 2
time ./word2phrase -train news.2012.en.shuffled-norm0-phrase0 -output
  ↪    news.2012.en.shuffled-norm0-phrase1 -threshold 100 -debug 2
tr A-Z a-z < news.2012.en.shuffled-norm0-phrase1 > news.2012.en.shuffled-norm1-phrase1
## Model...
time ./word2vec -train news.2012.en.shuffled-norm1-phrase1 -output vectors-phrase.bin -cbow 1 -size
  ↪    200 -window 10 -negative 25 -hs 0 -sample 1e-5 -threads 20 -binary 1 -iter 15
## Deploy...
./distance vectors-phrase.bin
```

## demo-phrases.sh output (1/3)

```
and@and-dblab:~/Documents/Projects/C++/word2vec/trunk$ ./demo-phrases.sh
make: Nothing to be done for 'all'.
--2015-11-12 18:33:08--
  ↪    http://www.statmt.org/wmt14/training-monolingual-news-crawl/news.2012.en.shuffled.gz
Resolving www.statmt.org (www.statmt.org)... 129.215.197.100
Connecting to www.statmt.org (www.statmt.org)|129.215.197.100|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 786717767 (750M) [application/x-gzip]
Saving to: news.2012.en.shuffled.gz

news.2012.en.shuffled.gz
  ↪    100%[==============================================>] 750.27M  4.25MB/s   in 3m 7s

2015-11-12 18:36:16 (4.02 MB/s) - news.2012.en.shuffled.gz saved [786717767/786717767]

...
```

## demo-phrases.sh output (2/3)

```
...
Starting training using file news.2012.en.shuffled-norm0
Words processed: 296900K     Vocab size: 33198K
Vocab size (unigrams + bigrams): 18838711
Words in train file: 296901342
Words written: 296900K
real  7m38.607s
user  7m8.592s
sys  0m15.176s

Starting training using file news.2012.en.shuffled-norm0-phrase0
Words processed: 280500K     Vocab size: 38761K
Vocab size (unigrams + bigrams): 21728781
Words in train file: 280513979
Words written: 280500K
real  7m0.022s
user  6m19.436s
sys  0m14.756s
...
```

## demo-phrases.sh output (3/3)

```
...
Starting training using file news.2012.en.shuffled-norm1-phrase1
Vocab size: 681320
Words in train file: 283545447
Alpha: 0.000005  Progress: 100.00%  Words/thread/sec: 162.97k
real  115m6.531s
user  434m57.904s
sys  1m4.464s
Enter word or sentence (EXIT to break):
```

-