Clustering the stack traces of ML/DL applications using various models

Amin Ghadesi, Roozbeh Aghili

Winter 2022

Problem

- Bug triage
- A lot of bugs
- Without any label data

Importance

- Saves time and energy
- Correct assignment of errors
- Clustering not Classification

Problem

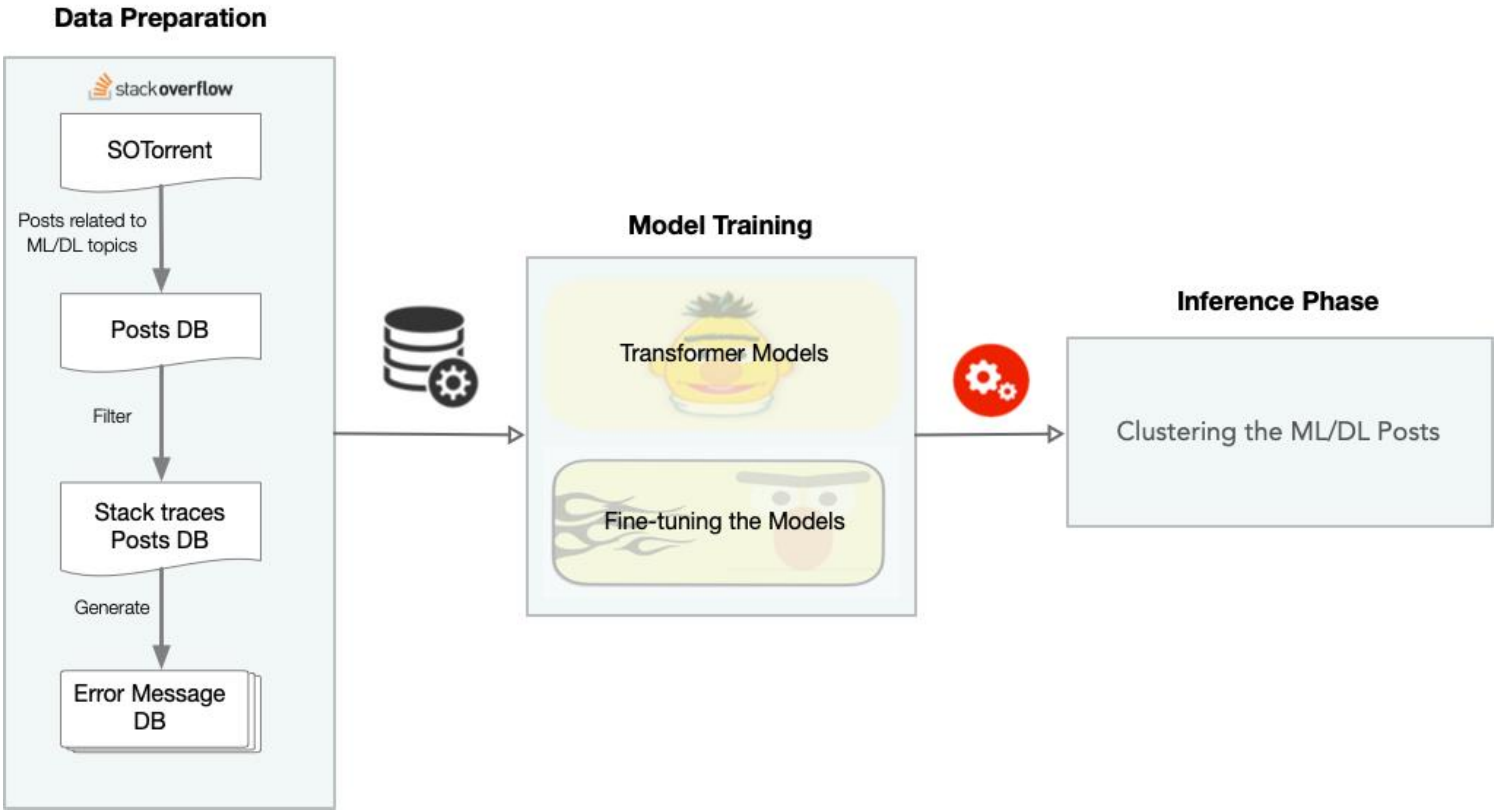
- Bug triage
- A lot of bugs
- Without any label data

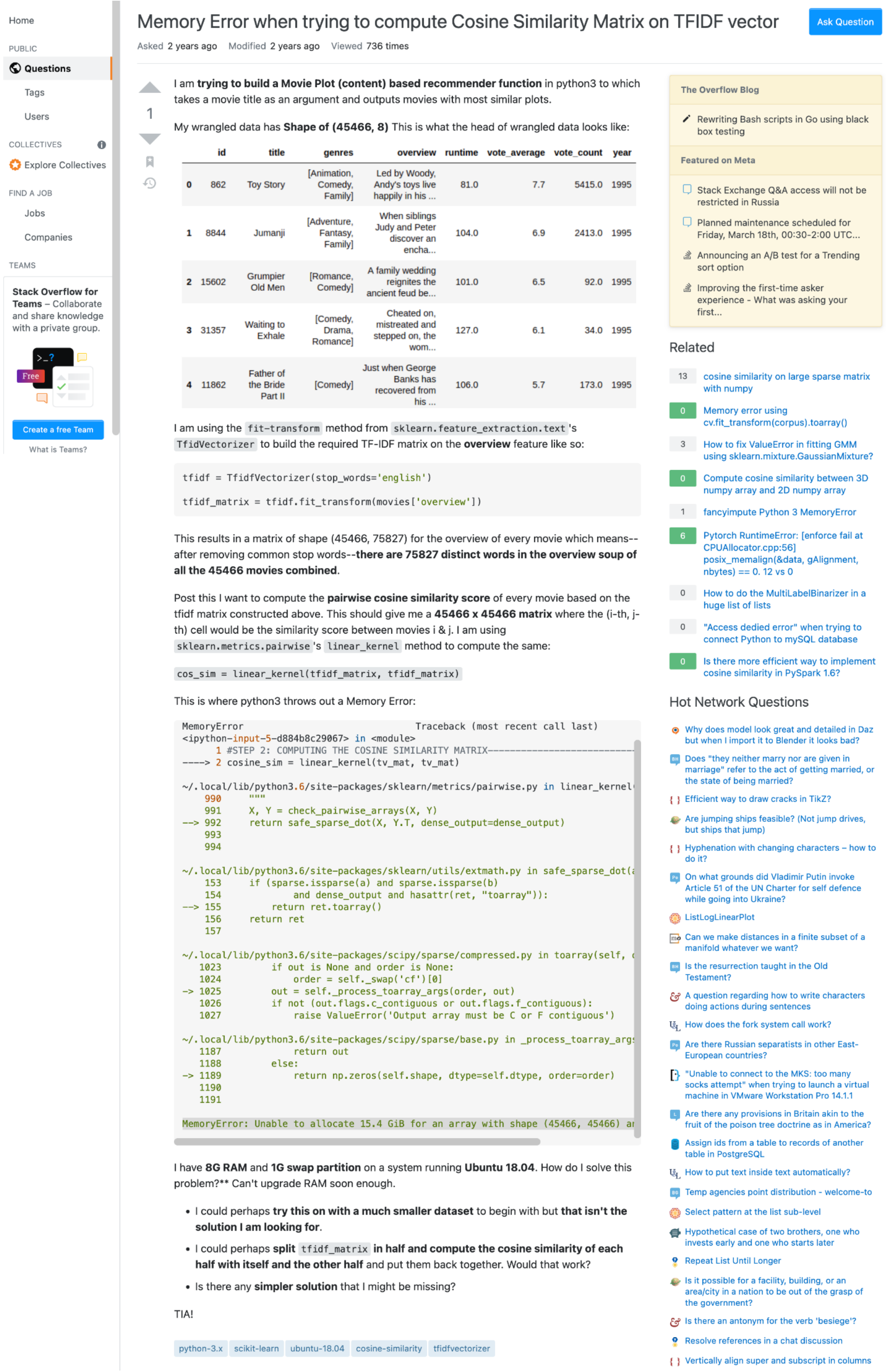
eclipse

337k bugs
2001 - 2010

Importance

- Saves time and energy
- Correct assignment of errors
- Clustering not Classification


Data Preparation
stackoverflow
SOTorrent
Posts related to ML/DL topics
Posts DB
Filter
Stack traces Posts DB
Generate
Error Message DB
Model Training
Transformer Models
Fine-tuning the Models
Inference Phase
Clustering the ML/DL Posts



Stack Overflow for Teams – Collaborate and share knowledge with a private group.

Memory Error when trying to compute Cosine Similarity Matrix on TFIDF vector

Asked 2 years ago Modified 2 years ago Viewed 736 times

I am **trying to build a Movie Plot (content) based recommender function** in python3 to which takes a movie title as an argument and outputs movies with most similar plots.

My wrangled data has **Shape of (45466, 8)** This is what the head of wrangled data looks like:

	id	title	genres	overview	runtime	vote_average	vote_count	year
0	862	Toy Story	[Animation, Comedy, Family]	Led by Woody, Andy's toys live happily in his ...	81.0	7.7	5415.0	1995
1	8844	Jumanji	[Adventure, Fantasy, Family]	When siblings Judy and Peter discover an encha...	104.0	6.9	2413.0	1995
2	15602	Grumpier Old Men	[Romance, Comedy]	A family wedding reignites the ancient feud be...	101.0	6.5	92.0	1995
3	31357	Waiting to Exhale	[Comedy, Drama, Romance]	Cheated on, mistreated and stepped on, the wom...	127.0	6.1	34.0	1995
4	11862	Father of the Bride Part II	[Comedy]	Just when George Banks has recovered from his ...	106.0	5.7	173.0	1995

I am using the `fit-transform` method from `sklearn.feature_extraction.text`'s `TfidVectorizer` to build the required TF-IDF matrix on the **overview** feature like so:

```
tfidf = TfidfVectorizer(stop_words='english')

tfidf_matrix = tfidf.fit_transform(movies['overview'])
```

This results in a matrix of shape (45466, 75827) for the overview of every movie which means--after removing common stop words--**there are 75827 distinct words in the overview soup of all the 45466 movies combined**.

Post this I want to compute the **pairwise cosine similarity score** of every movie based on the tfidf matrix constructed above. This should give me a **45466 x 45466 matrix** where the (i-th, j-th) cell would be the similarity score between movies i & j. I am using `sklearn.metrics.pairwise`'s `linear_kernel` method to compute the same:

`cos_sim = linear_kernel(tfidf_matrix, tfidf_matrix)`

This is where python3 throws out a Memory Error:

```
MemoryError                               Traceback (most recent call last)
<ipython-input-5-d884b8c29067> in <module>
      1 #STEP 2: COMPUTING THE COSINE SIMILARITY MATRIX--------------------------
----> 2 cosine_sim = linear_kernel(tv_mat, tv_mat)

~/.local/lib/python3.6/site-packages/sklearn/metrics/pairwise.py in linear_kernel
    990     """
    991     X, Y = check_pairwise_arrays(X, Y)
--> 992     return safe_sparse_dot(X, Y.T, dense_output=dense_output)
    993
    994

~/.local/lib/python3.6/site-packages/sklearn/utils/extmath.py in safe_sparse_dot(a
    153     if (sparse.issparse(a) and sparse.issparse(b)
    154             and dense_output and hasattr(ret, "toarray")):
--> 155         return ret.toarray()
    156     return ret
    157

~/.local/lib/python3.6/site-packages/scipy/sparse/compressed.py in toarray(self, c
   1023         if out is None and order is None:
   1024             order = self._swap('cf')[0]
-> 1025         out = self._process_toarray_args(order, out)
   1026         if not (out.flags.c_contiguous or out.flags.f_contiguous):
   1027             raise ValueError('Output array must be C or F contiguous')

~/.local/lib/python3.6/site-packages/scipy/sparse/base.py in _process_toarray_args
   1187             return out
   1188         else:
-> 1189             return np.zeros(self.shape, dtype=self.dtype, order=order)
   1190
   1191

MemoryError: Unable to allocate 15.4 GiB for an array with shape (45466, 45466) ar
```

I have **8G RAM** and **1G swap partition** on a system running **Ubuntu 18.04**. How do I solve this problem?** Can't upgrade RAM soon enough.

- I could perhaps **try this on with a much smaller dataset** to begin with but **that isn't the solution I am looking for**.
- I could perhaps **split** `tfidf_matrix` **in half and compute the cosine similarity of each half with itself and the other half** and put them back together. Would that work?
- Is there any **simpler solution** that I might be missing?

TIA!

python-3.x scikit-learn ubuntu-18.04 cosine-similarity tfidfvectorizer

The Overflow Blog

- Rewriting Bash scripts in Go using black box testing

Featured on Meta

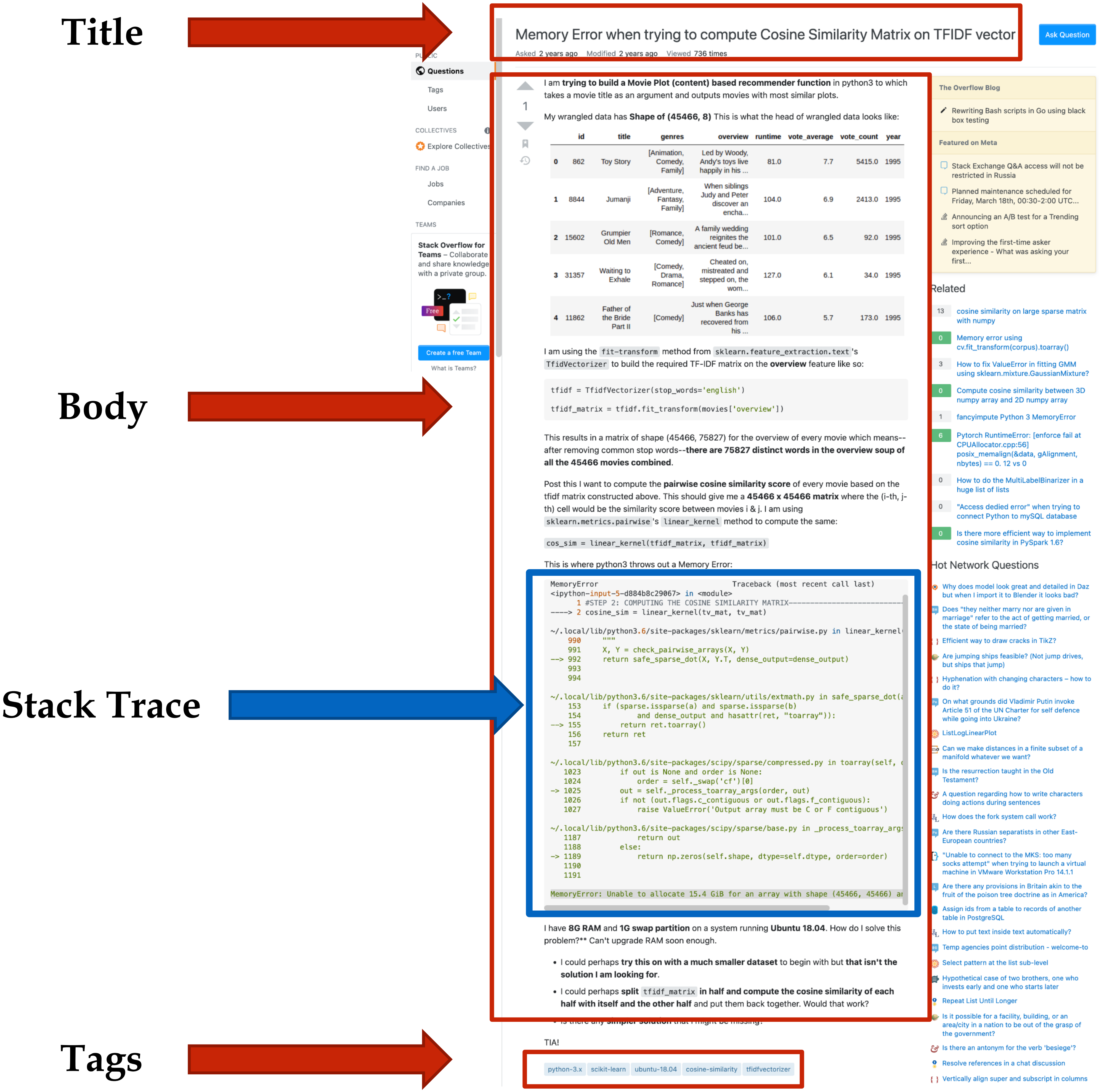
- Stack Exchange Q&A access will not be restricted in Russia
- Planned maintenance scheduled for Friday, March 18th, 00:30-2:00 UTC...
- Announcing an A/B test for a Trending sort option
- Improving the first-time asker experience - What was asking your first...

Related

13	cosine similarity on large sparse matrix with numpy
0	Memory error using cv.fit_transform(corpus).toarray()
3	How to fix ValueError in fitting GMM using sklearn.mixture.GaussianMixture?
0	Compute cosine similarity between 3D numpy array and 2D numpy array
1	fancyimpute Python 3 MemoryError
6	Pytorch RuntimeError: [enforce fail at CPUAllocator.cpp:56] posix_memalign(&data, gAlignment, nbytes) == 0. 12 vs 0
0	How to do the MultiLabelBinarizer in a huge list of lists
0	"Access dedied error" when trying to connect Python to mySQL database
0	Is there more efficient way to implement cosine similarity in PySpark 1.6?

Hot Network Questions

- Why does model look great and detailed in Daz but when I import it to Blender it looks bad?
- Does "they neither marry nor are given in marriage" refer to the act of getting married, or the state of being married?
- Efficient way to draw cracks in TikZ?
- Are jumping ships feasible? (Not jump drives, but ships that jump)
- Hyphenation with changing characters – how to do it?
- On what grounds did Vladimir Putin invoke Article 51 of the UN Charter for self defence while going into Ukraine?
- ListLogLinearPlot
- Can we make distances in a finite subset of a manifold whatever we want?
- Is the resurrection taught in the Old Testament?
- A question regarding how to write characters doing actions during sentences
- How does the fork system call work?
- Are there Russian separatists in other East-European countries?
- "Unable to connect to the MKS: too many socks attempt" when trying to launch a virtual machine in VMware Workstation Pro 14.1.1
- Are there any provisions in Britain akin to the fruit of the poison tree doctrine as in America?
- Assign ids from a table to records of another table in PostgreSQL
- How to put text inside text automatically?
- Temp agencies point distribution - welcome-to
- Select pattern at the list sub-level
- Hypothetical case of two brothers, one who invests early and one who starts later
- Repeat List Until Longer
- Is it possible for a facility, building, or an area/city in a nation to be out of the grasp of the government?
- Is there an antonym for the verb 'besiege'?
- Resolve references in a chat discussion
- Vertically align super and subscript in columns



Title

Body

Stack Trace

Tags

Memory Error when trying to compute Cosine Similarity Matrix on TFIDF vector

Asked 2 years ago Modified 2 years ago Viewed 736 times

Stack Overflow for Teams – Collaborate and share knowledge with a private group.

I am **trying to build a Movie Plot (content) based recommender function** in python3 to which takes a movie title as an argument and outputs movies with most similar plots.

My wrangled data has **Shape of (45466, 8)** This is what the head of wrangled data looks like:

	id	title	genres	overview	runtime	vote_average	vote_count	year
0	862	Toy Story	[Animation, Comedy, Family]	Led by Woody, Andy's toys live happily in his ...	81.0	7.7	5415.0	1995
1	8844	Jumanji	[Adventure, Fantasy, Family]	When siblings Judy and Peter discover an encha...	104.0	6.9	2413.0	1995
2	15602	Grumpier Old Men	[Romance, Comedy]	A family wedding reignites the ancient feud be...	101.0	6.5	92.0	1995
3	31357	Waiting to Exhale	[Comedy, Drama, Romance]	Cheated on, mistreated and stepped on, the wom...	127.0	6.1	34.0	1995
4	11862	Father of the Bride Part II	[Comedy]	Just when George Banks has recovered from his ...	106.0	5.7	173.0	1995

I am using the `fit-transform` method from `sklearn.feature_extraction.text`'s `TfidVectorizer` to build the required TF-IDF matrix on the **overview** feature like so:

```
tfidf = TfidfVectorizer(stop_words='english')

tfidf_matrix = tfidf.fit_transform(movies['overview'])
```

This results in a matrix of shape (45466, 75827) for the overview of every movie which means--after removing common stop words--**there are 75827 distinct words in the overview soup of all the 45466 movies combined**.

Post this I want to compute the **pairwise cosine similarity score** of every movie based on the tfidf matrix constructed above. This should give me a **45466 x 45466 matrix** where the (i-th, j-th) cell would be the similarity score between movies i & j. I am using `sklearn.metrics.pairwise`'s `linear_kernel` method to compute the same:

`cos_sim = linear_kernel(tfidf_matrix, tfidf_matrix)`

This is where python3 throws out a Memory Error:

```
MemoryError                               Traceback (most recent call last)
<ipython-input-5-d884b8c29067> in <module>
      1 #STEP 2: COMPUTING THE COSINE SIMILARITY MATRIX--------------------------
----> 2 cosine_sim = linear_kernel(tv_mat, tv_mat)

~/.local/lib/python3.6/site-packages/sklearn/metrics/pairwise.py in linear_kernel(
    990     """
    991     X, Y = check_pairwise_arrays(X, Y)
--> 992     return safe_sparse_dot(X, Y.T, dense_output=dense_output)
    993
    994

~/.local/lib/python3.6/site-packages/sklearn/utils/extmath.py in safe_sparse_dot(a
    153     if (sparse.issparse(a) and sparse.issparse(b)
    154             and dense_output and hasattr(ret, "toarray")):
--> 155         return ret.toarray()
    156     return ret
    157

~/.local/lib/python3.6/site-packages/scipy/sparse/compressed.py in toarray(self, o
   1023         if out is None and order is None:
   1024             order = self._swap('cf')[0]
-> 1025         out = self._process_toarray_args(order, out)
   1026         if not (out.flags.c_contiguous or out.flags.f_contiguous):
   1027             raise ValueError('Output array must be C or F contiguous')

~/.local/lib/python3.6/site-packages/scipy/sparse/base.py in _process_toarray_args
   1187             return out
   1188         else:
-> 1189             return np.zeros(self.shape, dtype=self.dtype, order=order)
   1190
   1191

MemoryError: Unable to allocate 15.4 GiB for an array with shape (45466, 45466) ar
```

I have **8G RAM** and **1G swap partition** on a system running **Ubuntu 18.04**. How do I solve this problem?** Can't upgrade RAM soon enough.

- I could perhaps **try this on with a much smaller dataset** to begin with but **that isn't the solution I am looking for**.
- I could perhaps **split** `tfidf_matrix` **in half and compute the cosine similarity of each half with itself and the other half** and put them back together. Would that work?
- Is there any **simpler solution** that I might be missing?

TIA!

python-3.x scikit-learn ubuntu-18.04 cosine-similarity tfidfvectorizer

The Overflow Blog

- Rewriting Bash scripts in Go using black box testing

Featured on Meta

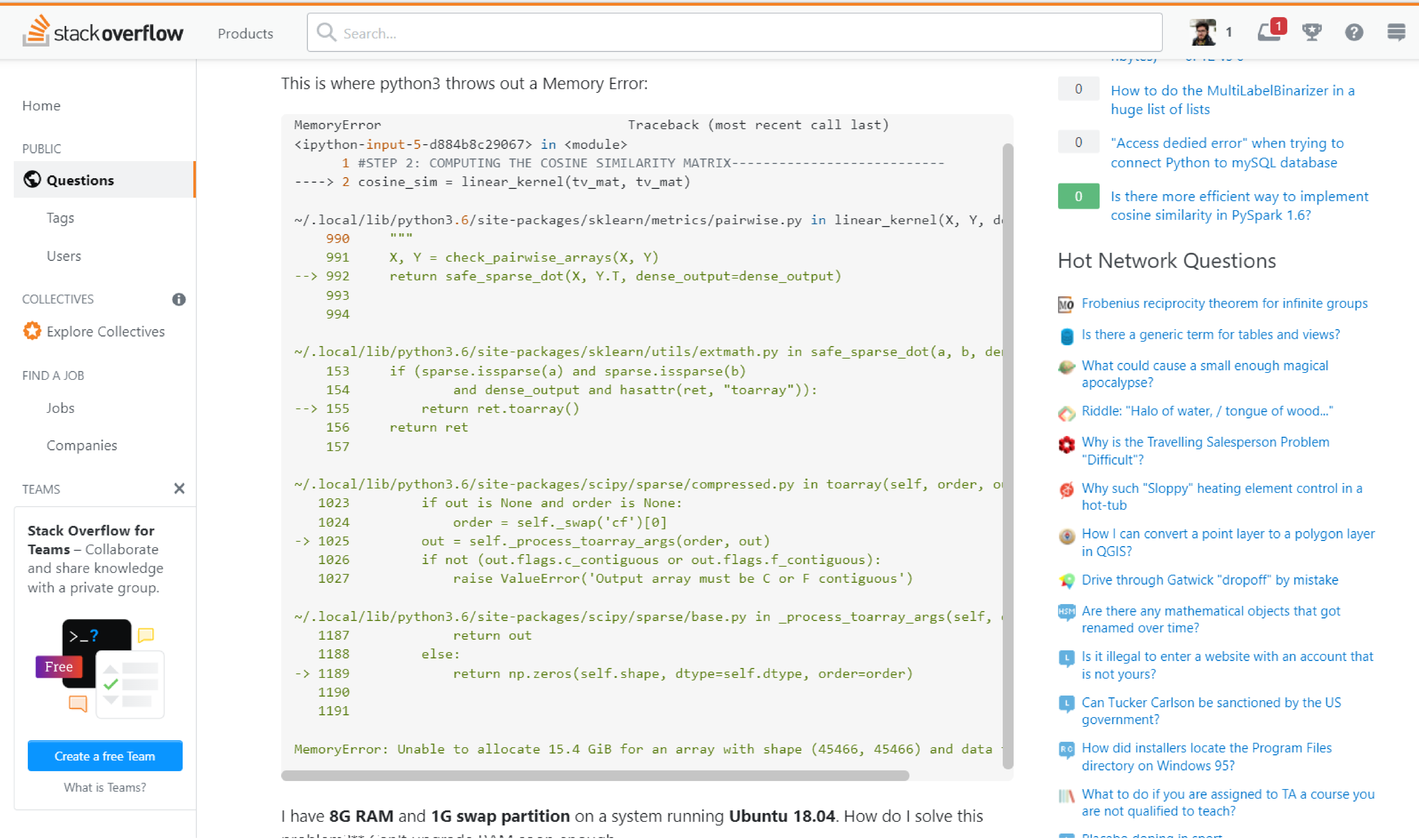
- Stack Exchange Q&A access will not be restricted in Russia
- Planned maintenance scheduled for Friday, March 18th, 00:30-2:00 UTC...
- Announcing an A/B test for a Trending sort option
- Improving the first-time asker experience - What was asking your first...

Related

13	cosine similarity on large sparse matrix with numpy
0	Memory error using cv.fit_transform(corpus).toarray()
3	How to fix ValueError in fitting GMM using sklearn.mixture.GaussianMixture?
0	Compute cosine similarity between 3D numpy array and 2D numpy array
1	fancyimpute Python 3 MemoryError
6	Pytorch RuntimeError: [enforce fail at CPUAllocator.cpp:56] posix_memalign(&data, gAlignment, nbytes) == 0. 12 vs 0
0	How to do the MultiLabelBinarizer in a huge list of lists
0	"Access dedied error" when trying to connect Python to mySQL database
0	Is there more efficient way to implement cosine similarity in PySpark 1.6?

Hot Network Questions

- Why does model look great and detailed in Daz but when I import it to Blender it looks bad?
- Does "they neither marry nor are given in marriage" refer to the act of getting married, or the state of being married?
- Efficient way to draw cracks in TikZ?
- Are jumping ships feasible? (Not jump drives, but ships that jump)
- Hyphenation with changing characters – how to do it?
- On what grounds did Vladimir Putin invoke Article 51 of the UN Charter for self defence while going into Ukraine?
- ListLogLinearPlot
- Can we make distances in a finite subset of a manifold whatever we want?
- Is the resurrection taught in the Old Testament?
- A question regarding how to write characters doing actions during sentences
- How does the fork system call work?
- Are there Russian separatists in other East-European countries?
- "Unable to connect to the MKS: too many socks attempt" when trying to launch a virtual machine in VMware Workstation Pro 14.1.1
- Are there any provisions in Britain akin to the fruit of the poison tree doctrine as in America?
- Assign ids from a table to records of another table in PostgreSQL
- How to put text inside text automatically?
- Temp agencies point distribution - welcome-to
- Select pattern at the list sub-level
- Hypothetical case of two brothers, one who invests early and one who starts later
- Repeat List Until Longer
- Is it possible for a facility, building, or an area/city in a nation to be out of the grasp of the government?
- Is there an antonym for the verb 'besiege'?
- Resolve references in a chat discussion
- Vertically align super and subscript in columns



Stack Overflow for Teams – Collaborate and share knowledge with a private group.

This is where python3 throws out a Memory Error:

```
MemoryError                               Traceback (most recent call last)
<ipython-input-5-d884b8c29067> in <module>
      1 #STEP 2: COMPUTING THE COSINE SIMILARITY MATRIX--------------------------
----> 2 cosine_sim = linear_kernel(tv_mat, tv_mat)

~/.local/lib/python3.6/site-packages/sklearn/metrics/pairwise.py in linear_kernel(X, Y, d
    990     """
    991     X, Y = check_pairwise_arrays(X, Y)
--> 992     return safe_sparse_dot(X, Y.T, dense_output=dense_output)
    993 
    994 

~/.local/lib/python3.6/site-packages/sklearn/utils/extmath.py in safe_sparse_dot(a, b, de
    153     if (sparse.issparse(a) and sparse.issparse(b)
    154             and dense_output and hasattr(ret, "toarray")):
--> 155         return ret.toarray()
    156     return ret
    157 

~/.local/lib/python3.6/site-packages/scipy/sparse/compressed.py in toarray(self, order, o
   1023         if out is None and order is None:
   1024             order = self._swap('cf')[0]
-> 1025         out = self._process_toarray_args(order, out)
   1026         if not (out.flags.c_contiguous or out.flags.f_contiguous):
   1027             raise ValueError('Output array must be C or F contiguous')

~/.local/lib/python3.6/site-packages/scipy/sparse/base.py in _process_toarray_args(self, 
   1187             return out
   1188         else:
-> 1189             return np.zeros(self.shape, dtype=self.dtype, order=order)
   1190 
   1191 

MemoryError: Unable to allocate 15.4 GiB for an array with shape (45466, 45466) and data 
```

I have **8G RAM** and **1G swap partition** on a system running **Ubuntu 18.04**. How do I solve this

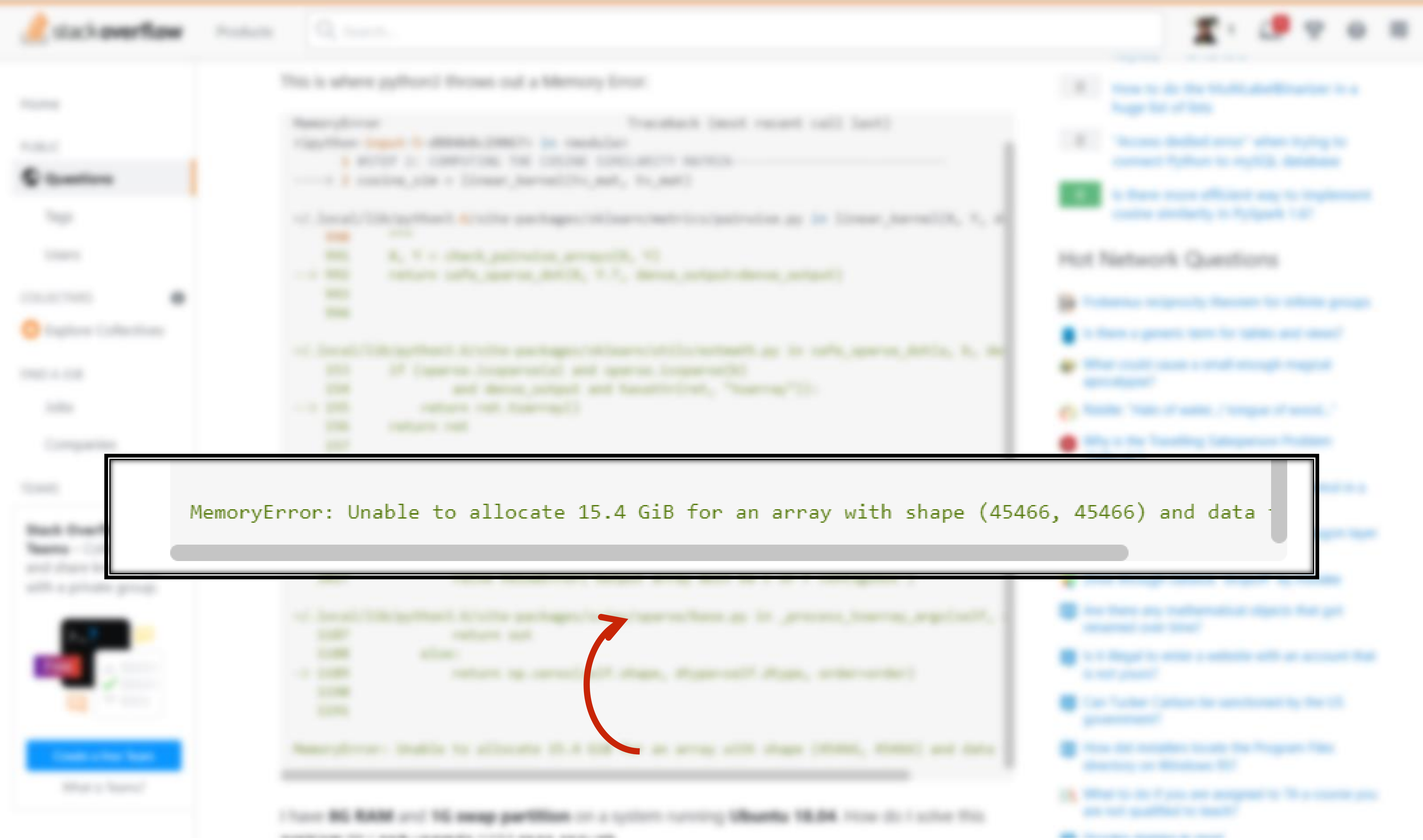
0 How to do the MultiLabelBinarizer in a huge list of lists

0 "Access dedied error" when trying to connect Python to mySQL database

0 Is there more efficient way to implement cosine similarity in PySpark 1.6?

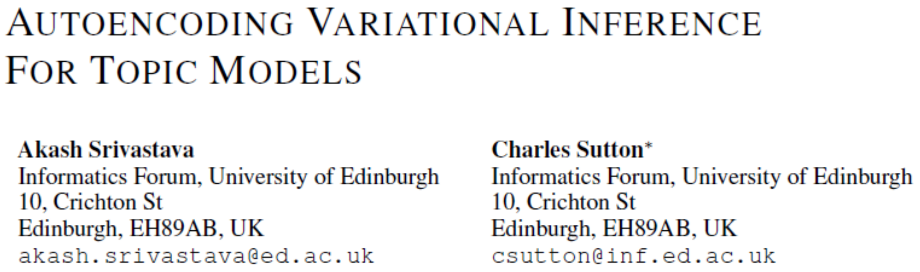
Hot Network Questions

- Frobenius reciprocity theorem for infinite groups
- Is there a generic term for tables and views?
- What could cause a small enough magical apocalypse?
- Riddle: "Halo of water, / tongue of wood..."
- Why is the Travelling Salesperson Problem "Difficult"?
- Why such "Sloppy" heating element control in a hot-tub
- How I can convert a point layer to a polygon layer in QGIS?
- Drive through Gatwick "dropoff" by mistake
- Are there any mathematical objects that got renamed over time?
- Is it illegal to enter a website with an account that is not yours?
- Can Tucker Carlson be sanctioned by the US government?
- How did installers locate the Program Files directory on Windows 95?
- What to do if you are assigned to TA a course you are not qualified to teach?
- Placebo doping in sport



```
MemoryError: Unable to allocate 15.4 GiB for an array with shape (45466, 45466) and data
```

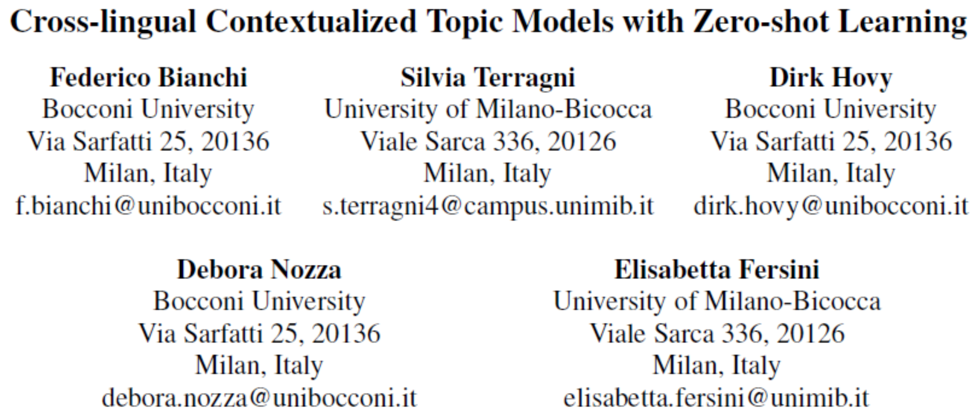
BERT
CodeBERT
ETM
NeuralLDA
CTM
LDA



AUTOENCODING VARIATIONAL INFERENCE
FOR TOPIC MODELS

Akash Srivastava
Informatics Forum, University of Edinburgh
10, Crichton St
Edinburgh, EH89AB, UK
akash.srivastava@ed.ac.uk

Charles Sutton*
Informatics Forum, University of Edinburgh
10, Crichton St
Edinburgh, EH89AB, UK
csutton@inf.ed.ac.uk



Cross-lingual Contextualized Topic Models with Zero-shot Learning

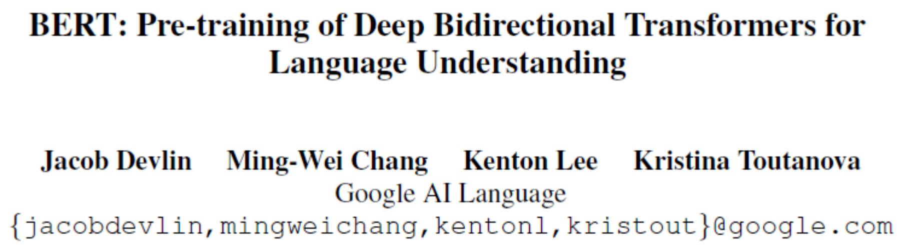
Federico Bianchi
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
f.bianchi@unibocconi.it

Silvia Terragni
University of Milano-Bicocca
Viale Sarca 336, 20126
Milan, Italy
s.terragni4@campus.unimib.it

Dirk Hovy
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
dirk.hovy@unibocconi.it

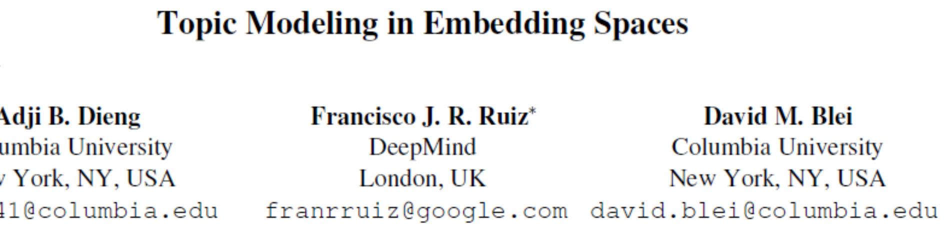
Debora Nozza
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
debora.nozza@unibocconi.it

Elisabetta Fersini
University of Milano-Bicocca
Viale Sarca 336, 20126
Milan, Italy
elisabetta.fersini@unimib.it



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

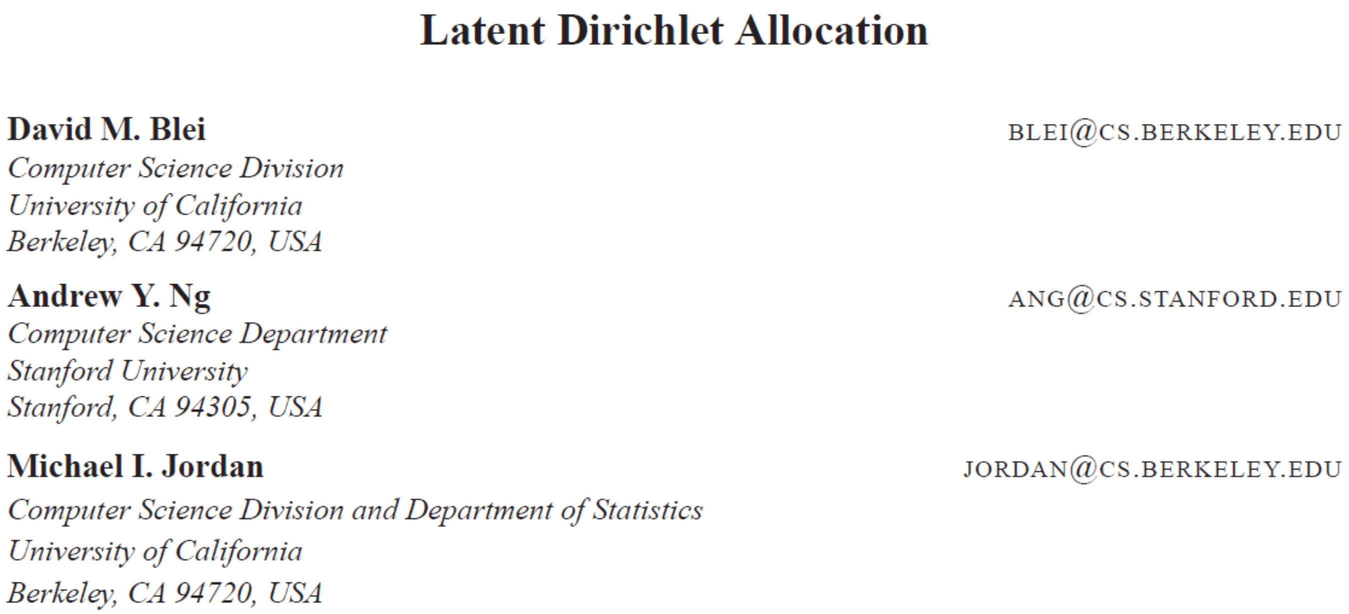


Topic Modeling in Embedding Spaces

Adji B. Dieng
Columbia University
New York, NY, USA
abd2141@columbia.edu

Francisco J. R. Ruiz*
DeepMind
London, UK
franrruiz@google.com

David M. Blei
Columbia University
New York, NY, USA
david.blei@columbia.edu

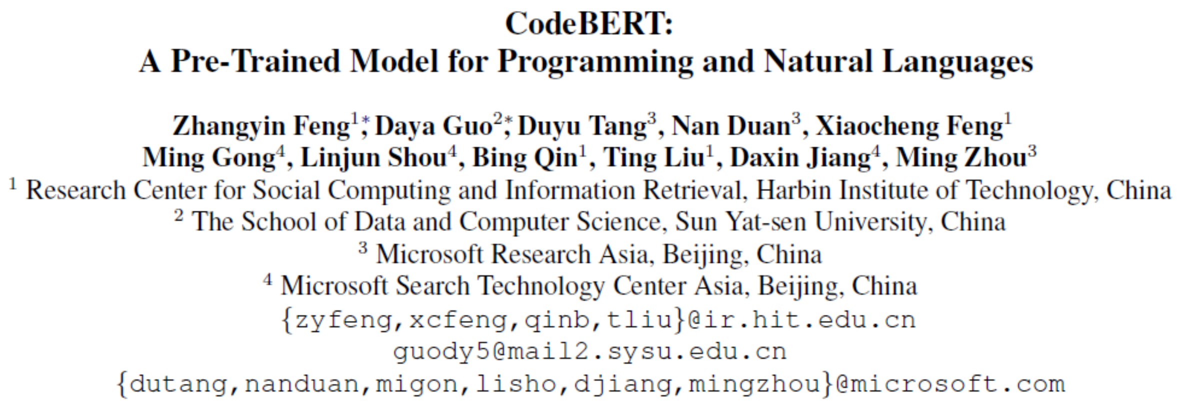


Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA



CodeBERT:
A Pre-Trained Model for Programming and Natural Languages

Zhangyin Feng[1]*, Daya Guo[2]*, Duyu Tang[3], Nan Duan[3], Xiaocheng Feng[1]
Ming Gong[4], Linjun Shou[4], Bing Qin[1], Ting Liu[1], Daxin Jiang[4], Ming Zhou[3]
[1] Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
[2] The School of Data and Computer Science, Sun Yat-sen University, China
[3] Microsoft Research Asia, Beijing, China
[4] Microsoft Search Technology Center Asia, Beijing, China
{zyfeng,xcfeng,qinb,tliu}@ir.hit.edu.cn
guody5@mail2.sysu.edu.cn
{dutang,nanduan,migon,lisho,djiang,mingzhou}@microsoft.com

Problem

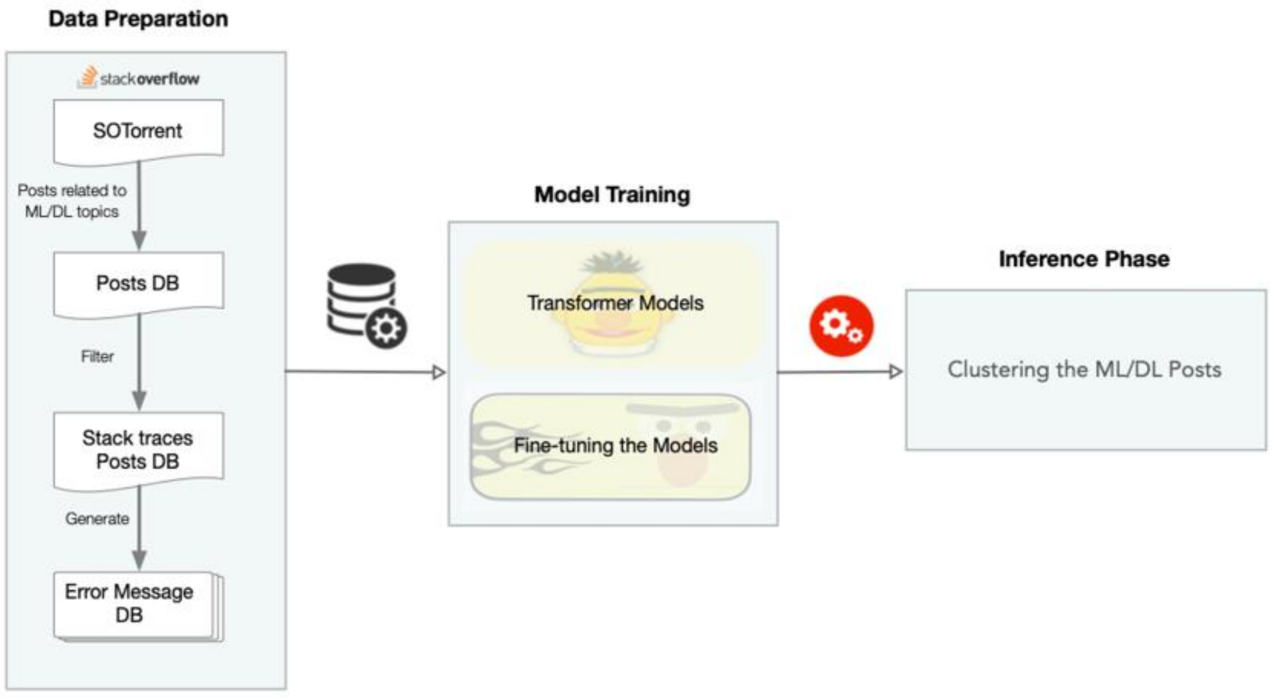
- Bug triage
- A lot of bugs
- Without any label data

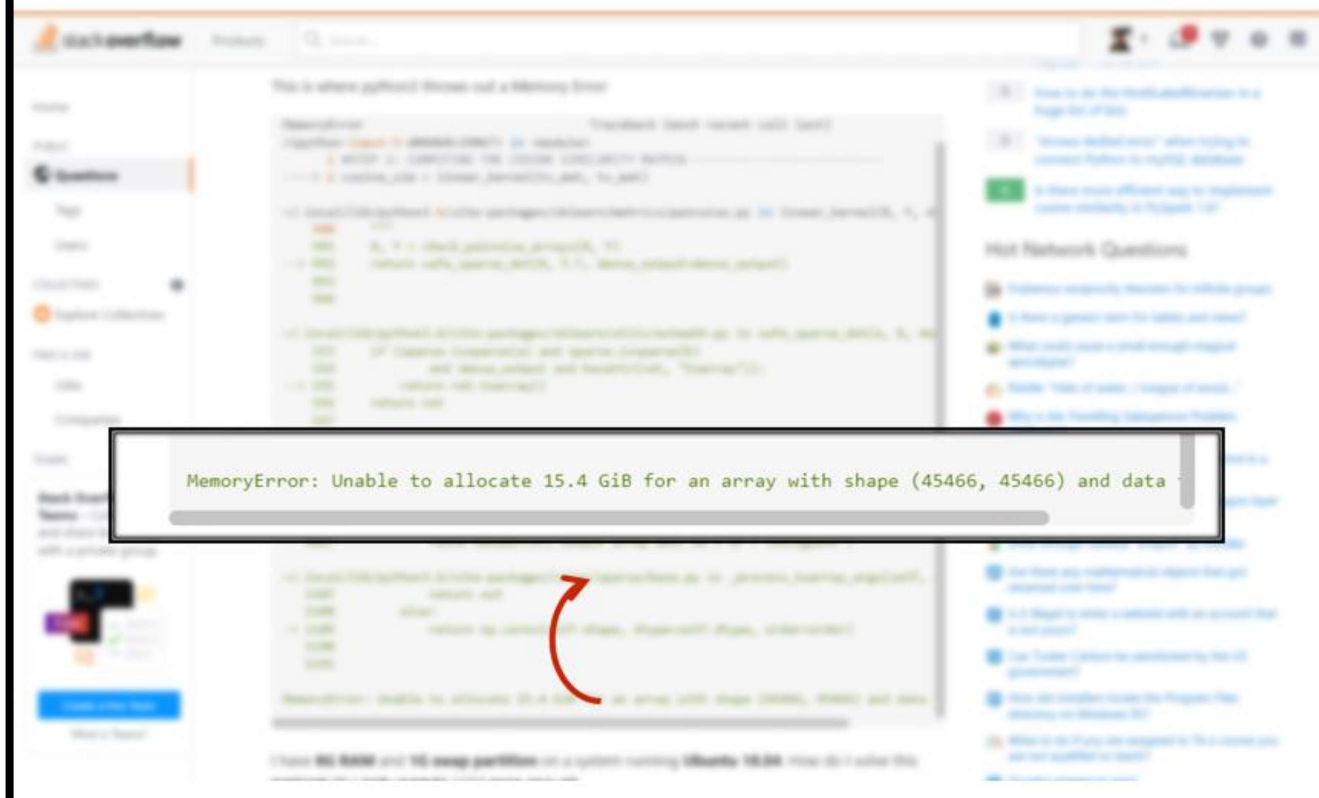
eclipse

337k bugs
2001 - 2010

Importance

- Saves time and energy
- Correct assignment of errors
- Clustering not Classification





```
MemoryError: Unable to allocate 15.4 GiB for an array with shape (45466, 45466) and data
```

BERT
CodeBERT
ETM
NeuralLDA
CTM
LDA

AUTOENCODING VARIATIONAL INFERENCE
FOR TOPIC MODELS

Akash Srivastava
Informatics Forum, University of Edinburgh
10, Crichton St
Edinburgh, EH89AB, UK
akash.srivastava@ed.ac.uk

Charles Sutton*
Informatics Forum, University of Edinburgh
10, Crichton St
Edinburgh, EH89AB, UK
csutton@inf.ed.ac.uk

Cross-lingual Contextualized Topic Models with Zero-shot Learning

Federico Bianchi
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
f.bianchi@unibocconi.it

Silvia Terragni
University of Milano-Bicocca
Viale Sarca 336, 20126
Milan, Italy
s.terragni4@campus.unimib.it

Dirk Hovy
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
dirk.hovy@unibocconi.it

Debora Nozza
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
debora.nozza@unibocconi.it

Elisabetta Fersini
University of Milano-Bicocca
Viale Sarca 336, 20126
Milan, Italy
elisabetta.fersini@unimib.it

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Topic Modeling in Embedding Spaces

Adji B. Dieng
Columbia University
New York, NY, USA
abd2141@columbia.edu

Francisco J. R. Ruiz*
DeepMind
London, UK
franrruiz@google.com

David M. Blei
Columbia University
New York, NY, USA
david.blei@columbia.edu

**CodeBERT:
A Pre-Trained Model for Programming and Natural Languages**

Zhangyin Feng[1]*, Daya Guo[2]*, Duyu Tang[3], Nan Duan[3], Xiaocheng Feng[1]
Ming Gong[4], Linjun Shou[4], Bing Qin[1], Ting Liu[1], Daxin Jiang[4], Ming Zhou[3]
[1] Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
[2] The School of Data and Computer Science, Sun Yat-sen University, China
[3] Microsoft Research Asia, Beijing, China
[4] Microsoft Search Technology Center Asia, Beijing, China
{zyfeng,xcfeng,qinb,tliu}@ir.hit.edu.cn
guody5@mail2.sysu.edu.cn
{dutang,nanduan,migon,lisho,djiang,mingzhou}@microsoft.com

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA