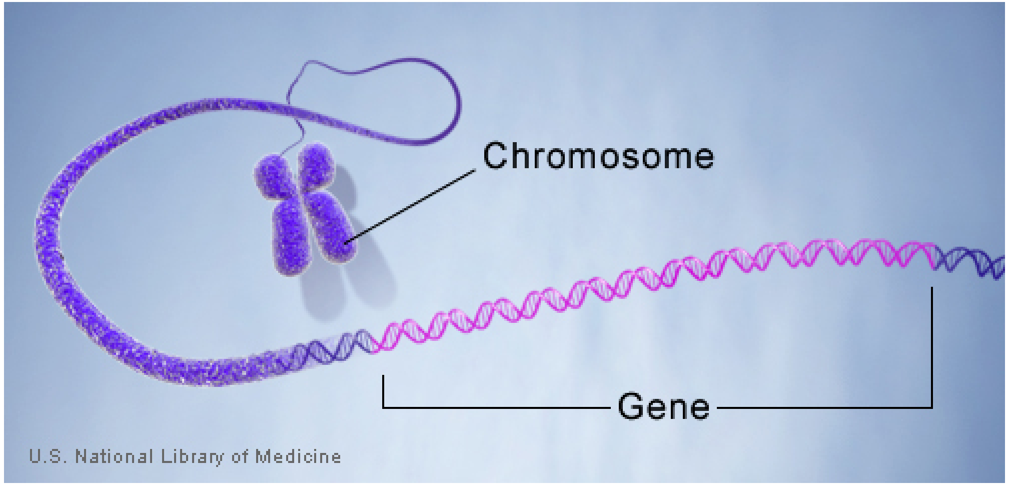
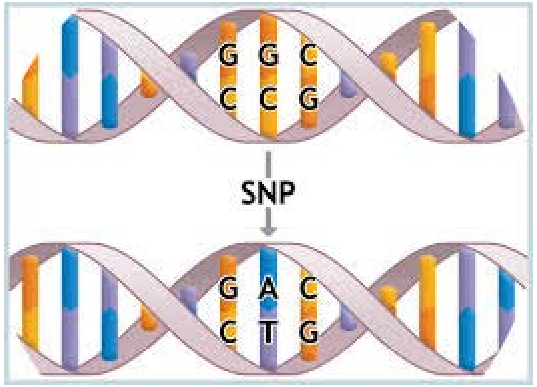
Masterthesis Philip Schledermann Knowledgebase construction of genetic variants in literature

Date: **2018-11-20**

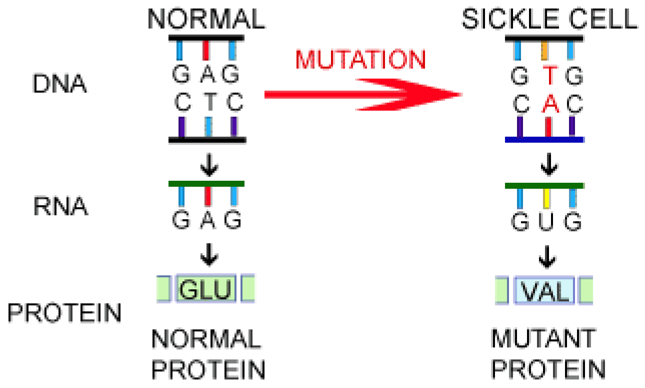
What are genes and mutations?





Left: „What is a gene?“ 18 July 2018 : https://ghr.nlm.nih.gov/primer/basics/gene
Rigth: „Single Nucleotide Polymorphism (SNP) Allele Frequency DNA Pools
„ 18 July 2018 : http://www.socmucimm.org/single-nucleotide-polymorphism-snp-allele-frequency-estimation-dna-po

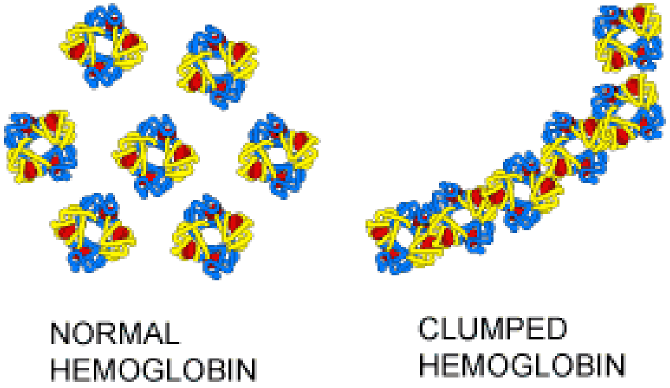
Consequences of a single change!



1 & 2 : Understanding Evolution. 2018. University of California Museum of Paleontology. 18 July 2018 <http://evolution.berkeley.edu/>.

3 : Sickle Cell Anemia- Types, Symptoms, Causes, Diagnosis and Treatment. 18 July 2018 <https://zovon.com/health-conditions/sickle-cell-anemia/>

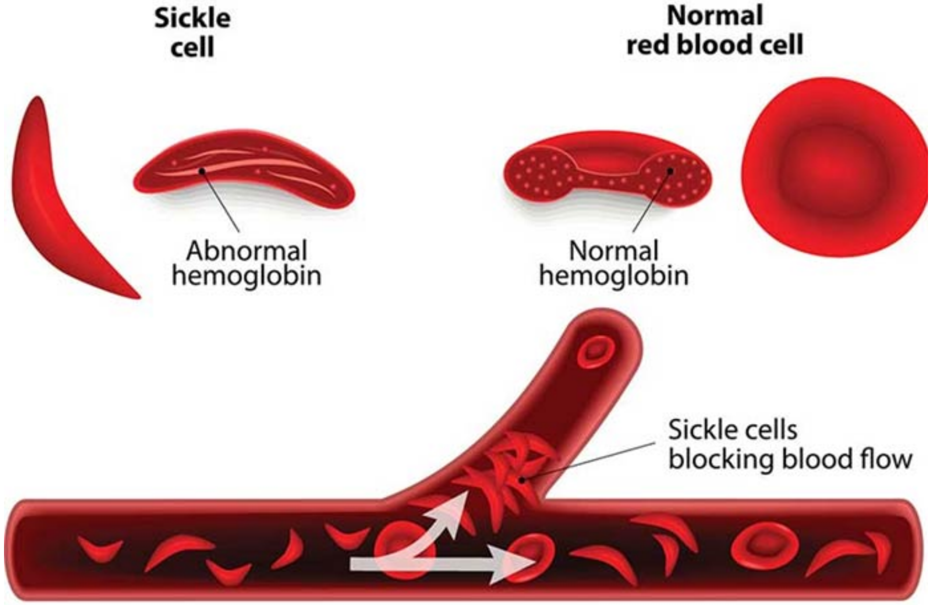
Consequences of a single change!



1 & 2 : Understanding Evolution. 2018. University of California Museum of Paleontology. 18 July 2018 <http://evolution.berkeley.edu/>.

3 : Sickle Cell Anemia- Types, Symptoms, Causes, Diagnosis and Treatment. 18 July 2018 <https://zovon.com/health-conditions/sickle-cell-anemia/>

Consequences of a single change!

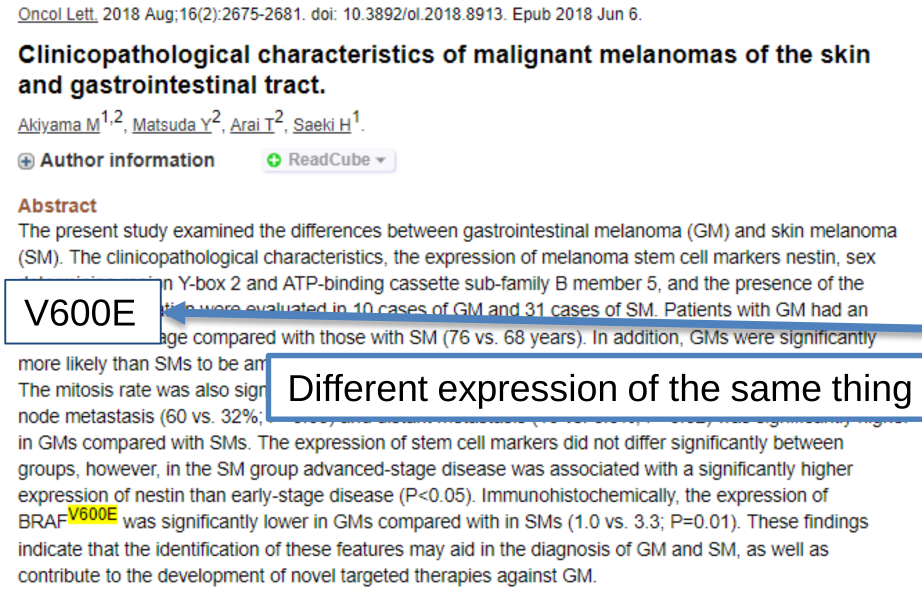


1 & 2 : Understanding Evolution. 2018. University of California Museum of Paleontology. 18 July 2018 <http://evolution.berkeley.edu/>.

3 : Sickle Cell Anemia- Types, Symptoms, Causes, Diagnosis and Treatment. 18 July 2018 <https://zovon.com/health-conditions/sickle-cell-anemia/>

Motivation

Different words, but same entity



Oncol Lett. 2018 Aug;16(2):2675-2681. doi: 10.3892/ol.2018.8913. Epub 2018 Jun 6.

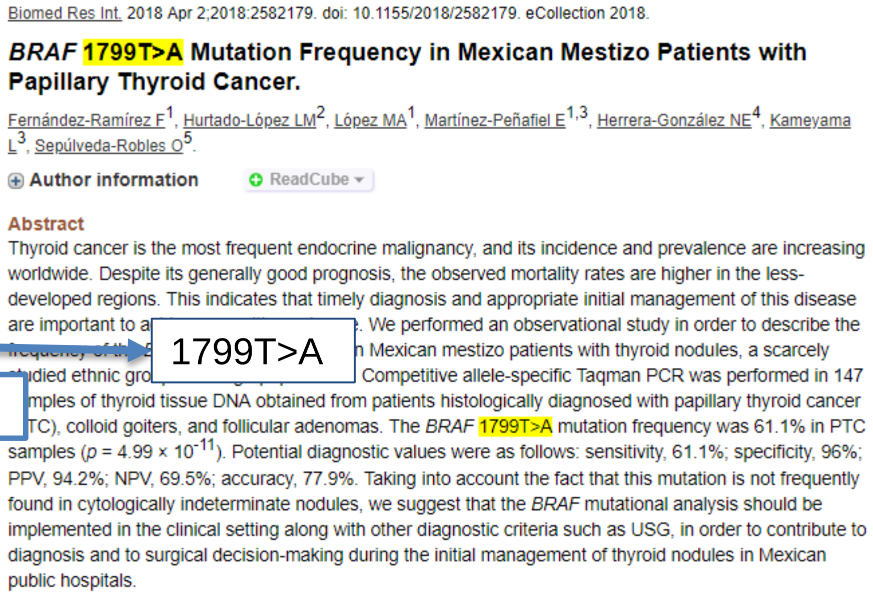
Clinicopathological characteristics of malignant melanomas of the skin and gastrointestinal tract.

Akiyama M[1,2], Matsuda Y[2], Arai T[2], Saeki H[1].

⊕ Author information ReadCube ▾

Abstract

The present study examined the differences between gastrointestinal melanoma (GM) and skin melanoma (SM). The clinicopathological characteristics, the expression of melanoma stem cell markers nestin, sex n Y-box 2 and ATP-binding cassette sub-family B member 5, and the presence of the were evaluated in 10 cases of GM and 31 cases of SM. Patients with GM had an age compared with those with SM (76 vs. 68 years). In addition, GMs were significantly more likely than SMs to be am The mitosis rate was also sign node metastasis (60 vs. 32%; in GMs compared with SMs. The expression of stem cell markers did not differ significantly between groups, however, in the SM group advanced-stage disease was associated with a significantly higher expression of nestin than early-stage disease (P<0.05). Immunohistochemically, the expression of BRAF$^{\text{V600E}}$ was significantly lower in GMs compared with in SMs (1.0 vs. 3.3; P=0.01). These findings indicate that the identification of these features may aid in the diagnosis of GM and SM, as well as contribute to the development of novel targeted therapies against GM.



Biomed Res Int. 2018 Apr 2;2018:2582179. doi: 10.1155/2018/2582179. eCollection 2018.

***BRAF* 1799T>A Mutation Frequency in Mexican Mestizo Patients with Papillary Thyroid Cancer.**

Fernández-Ramírez F[1], Hurtado-López LM[2], López MA[1], Martínez-Peñafiel E[1,3], Herrera-González NE[4], Kameyama L[3], Sepúlveda-Robles O[5].

⊕ Author information ReadCube ▾

Abstract

Thyroid cancer is the most frequent endocrine malignancy, and its incidence and prevalence are increasing worldwide. Despite its generally good prognosis, the observed mortality rates are higher in the less-developed regions. This indicates that timely diagnosis and appropriate initial management of this disease are important to a . We performed an observational study in order to describe the n Mexican mestizo patients with thyroid nodules, a scarcely studied ethnic gro Competitive allele-specific Taqman PCR was performed in 147 mples of thyroid tissue DNA obtained from patients histologically diagnosed with papillary thyroid cancer TC), colloid goiters, and follicular adenomas. The *BRAF* 1799T>A mutation frequency was 61.1% in PTC samples ($p = 4.99 \times 10^{-11}$). Potential diagnostic values were as follows: sensitivity, 61.1%; specificity, 96%; PPV, 94.2%; NPV, 69.5%; accuracy, 77.9%. Taking into account the fact that this mutation is not frequently found in cytologically indeterminate nodules, we suggest that the *BRAF* mutational analysis should be implemented in the clinical setting along with other diagnostic criteria such as USG, in order to contribute to diagnosis and to surgical decision-making during the initial management of thyroid nodules in Mexican public hospitals.

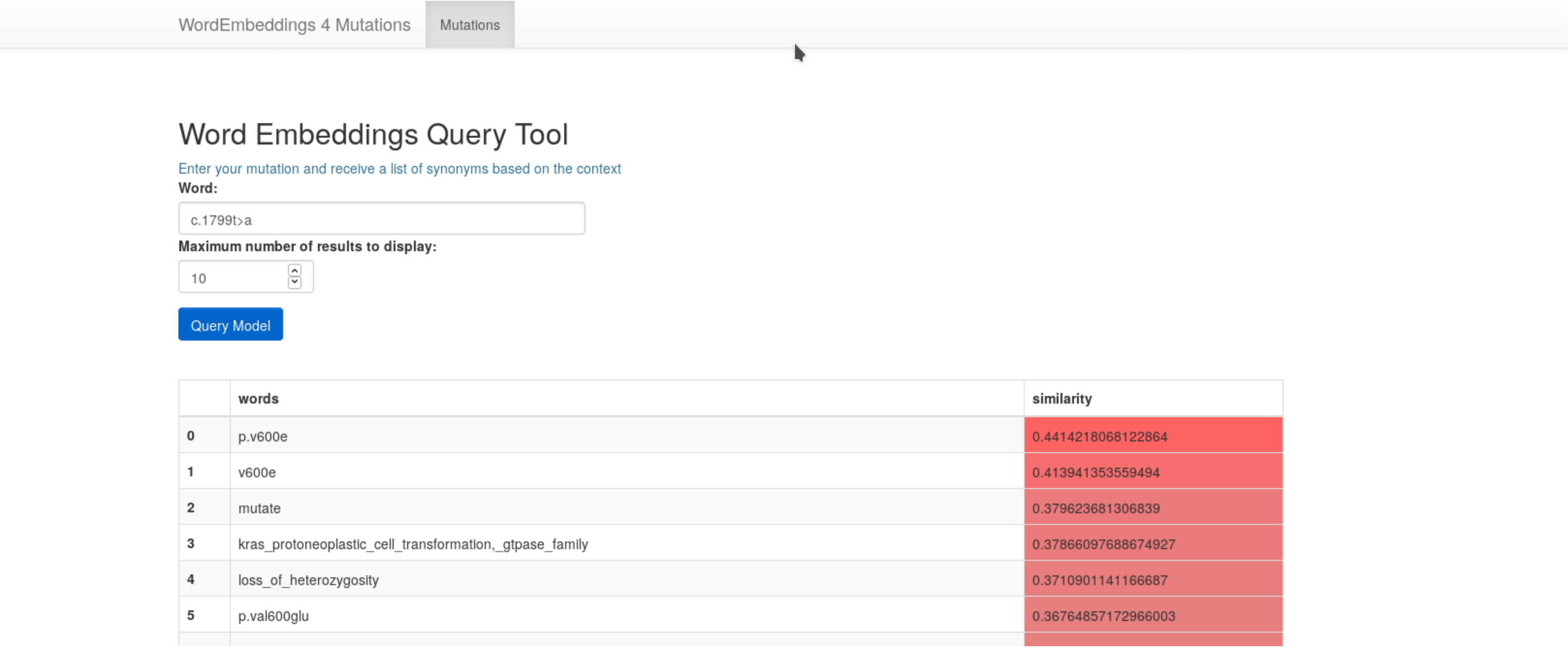
V600E

1799T>A

Different expression of the same thing

Left: 18 July 2018 PubMed: https://www.ncbi.nlm.nih.gov/pubmed/30013664
Right: 18 July 2018 PubMed : https://www.ncbi.nlm.nih.gov/pubmed/29808165

Knowledgebase of genetic variants and their synonyms



WordEmbeddings 4 Mutations

Mutations

Word Embeddings Query Tool

Enter your mutation and receive a list of synonyms based on the context

Word:

c.1799t>a

Maximum number of results to display:

10

Query Model

	words	similarity
0	p.v600e	0.4414218068122864
1	v600e	0.413941353559494
2	mutate	0.379623681306839
3	kras_protoneoplastic_cell_transformation,_gtpase_family	0.37866097688674927
4	loss_of_heterozygosity	0.3710901141166687
5	p.val600glu	0.36764857172966003

ClinVar
- human variation data

- Manual submission
- Manual curation

rs113488022

- c.1799T>A
- V600E
- g.140753336A>T
- p.Val600Glu
- g.176429T>A
- […]

Source: ClinVar 20181119 - https://www.ncbi.nlm.nih.gov/clinvar/variation/13961/

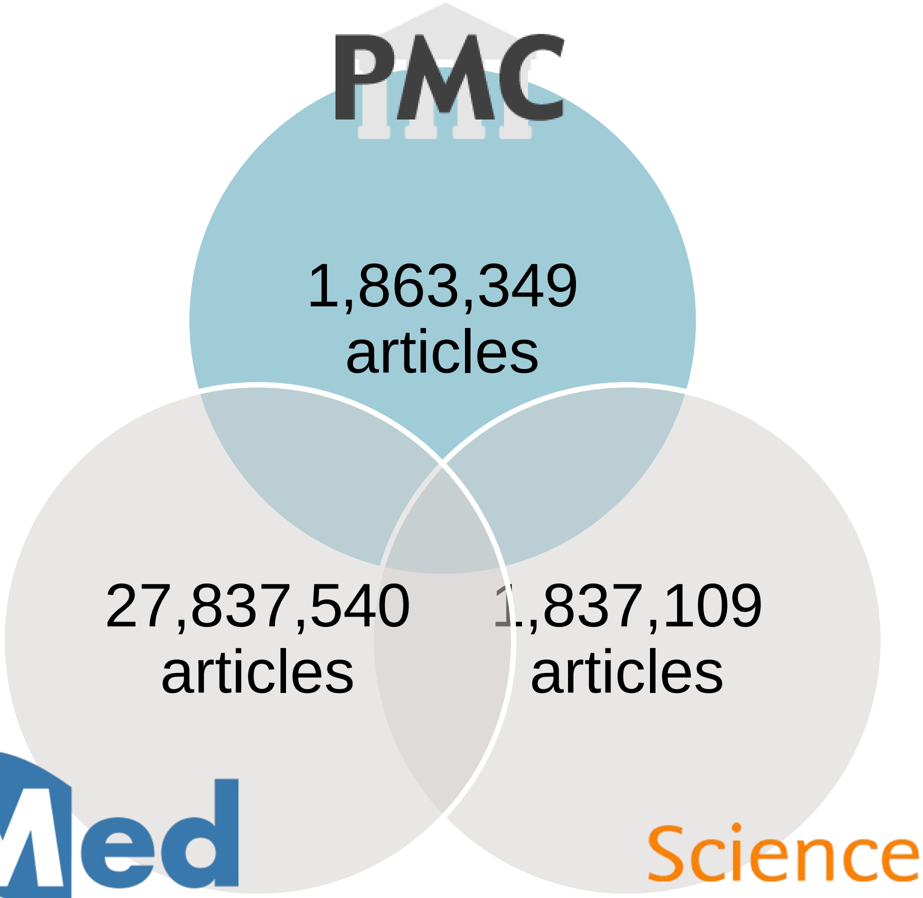
Goals

- Detection of rare mutation mentions by not writing rules
- Normalize/Link entities (dbSNP identifiers)
- Compare the usage of word embeddings for knowledge extraction

Methodology

1. Getting the text content out of the data
2. Creating two text corpora
   - Basic corpus
   - Cleansed corpus
3. Applying word embeddings on the words/tokens in the corpora
4. Evaluating the models against ClinVar (contains human variation data)

Source Data



PMC

1,863,349 articles

PubMed

27,837,540 articles

ScienceDirect

1,837,109 articles

Logos are trademarks and brands are property of their respecive owners

Two input sets for the models beeing created

Basic corpus

- Simple tokenization

Cleansed corpus

- Extensive cleansing and normalization applied where possible

Objectives with „Other Entities Tagged“

- Harmonize + simplify the text as much as possible
- Less tokens
  - Singular and plural forms are one
  - Removing stopwords
  - Eg. different company names, meaning the same entitiy are normalized to one word

better model

Replace other entities by the preferred label and do basic NLP

Input

"BRAF is not associated with non small cell lung carcinoma, but with c.1799A>T and V600E.“

Sentence in the **basic** corpus

“braf”, “is”, “not”, “associated”, “with”, “non”, “small”, “cell”, “lung”, “carcinoma”, “but”, “with”, “c.1799a>t”, “and”, “v600e”

Sentence in the **cleansed** corpus

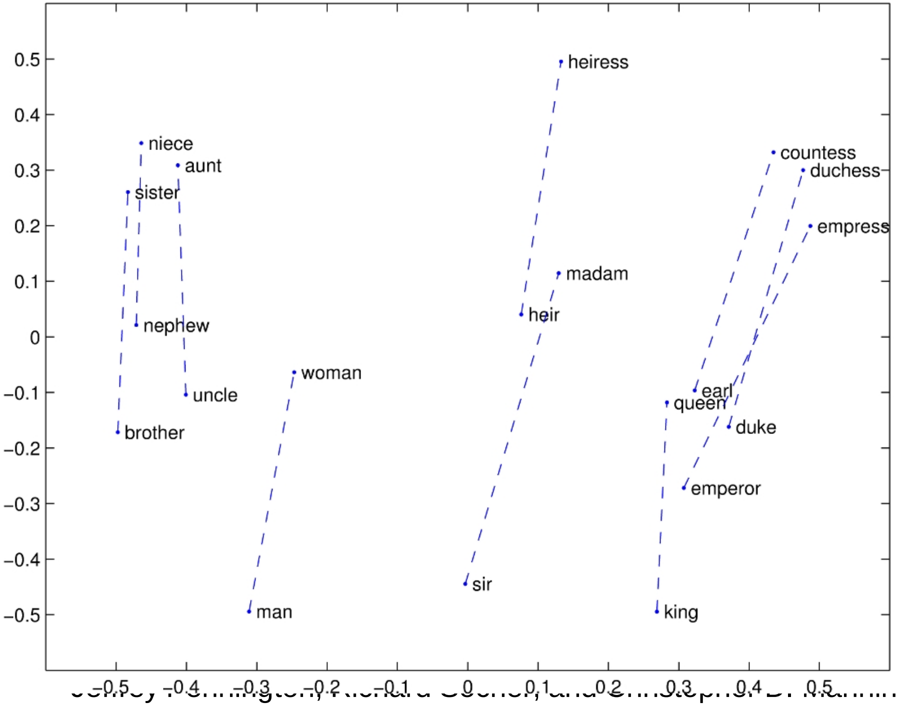
“b-raf_proto-oncogene,_serine/threonine_kinase”, “associate”, “non-small_cell_lung_cancer”, “c.1799a>t”, “v600e”

Firth, J. R. 1957:11

You shall know a word by the company it keeps

What are word embeddings?

- Training a shallow neural network that predicts the surroundings words
- Taking the hidden layer and intepreting it as word vectors
- Synonyms that share similar contexts are placed near each other
- Relations between contexts are preserved



Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

Corpus & model statistics

	Basic corpus	Cleansed corpus
Records	29,354,945	
Words	11,936,304,678	
Distinct words	18,790,878	
Distinct words after cleansing	18,790,878	52,057,405
Minimum count	25	40
Distinct words in model	1,317,892	1,398,581
Model construction runtime (shared HW)	11 hours	5 hours

Evaluation data

- Based on ClinVar
- V600E:
  - c.1799T>A
  - p.Val600Glu
  - rs113488022

Label

Synonyms

- With single letter and three-letter amino-acid codes, as well as with and without qualifier
- There are 350.832 records in the evaluation set

Evaluation – model only

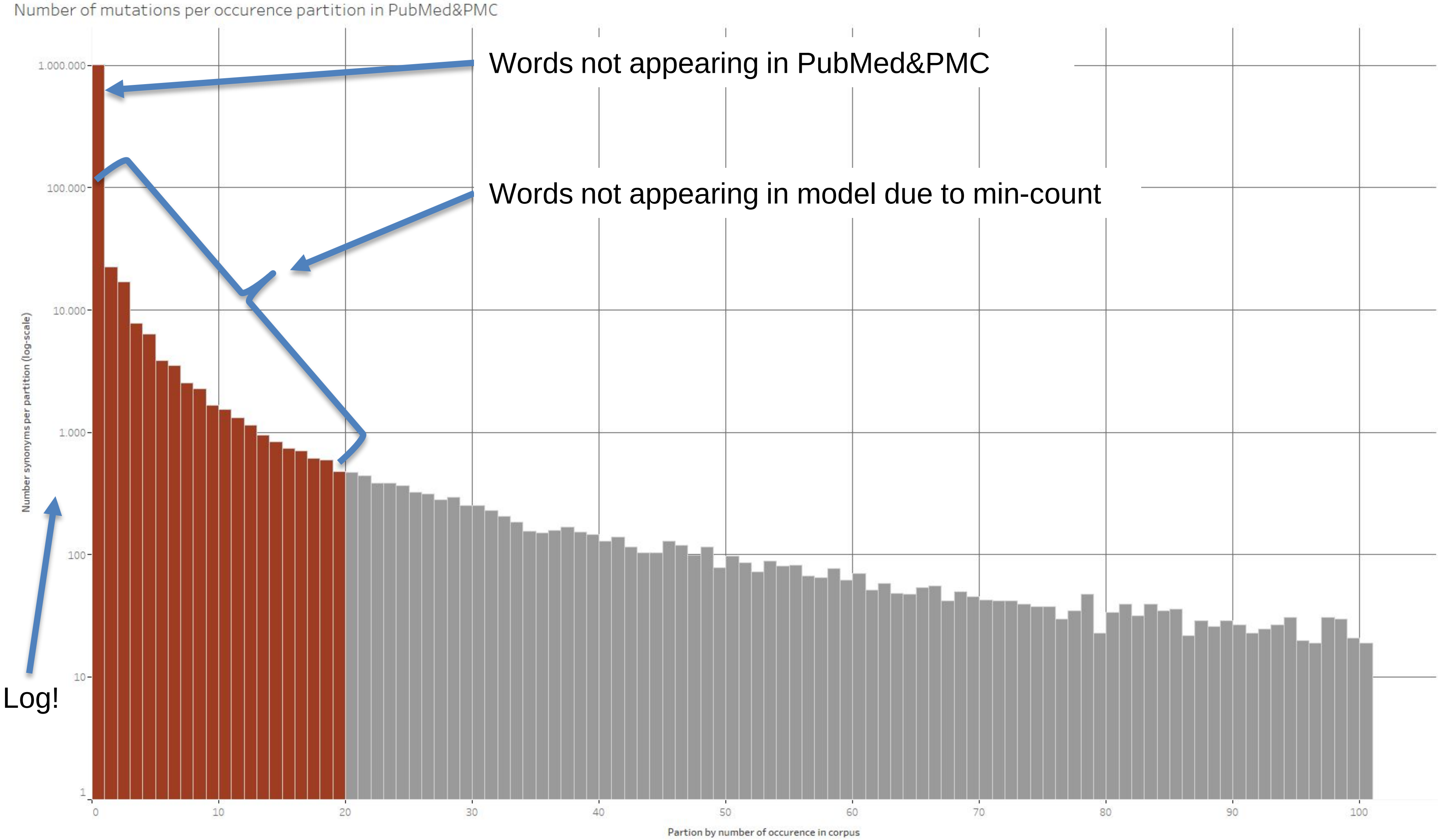
	Basic model		Cleansed model	
Question direction	Label>Syn	Syn>Label	Label>Syn	Syn>Label
Number of tests	1055	1041	217	287
Precision@K1	00.90%	01.63%	03.23%	02.32%
K1	2	1	1	3
Recall@K2	05.98%	11.71%	15.67%	23.88%
K2	117	120	107	119

Evaluation – model plus a mutation tagger

	Basic model		**Cleansed model**	
Question direction	Label>Syn	Syn>Label	Label>Syn	Syn>Label
Number of tests	86	125	42	71
Precision@K1	11.05%	13.60%	16.67%	09.67%
K1	1	1	1	1
Recall@K2	73.33%	97.52%	80.95%	96.53%
K2	120	120	107	119

Error analysis

- Not all false positives are false positives
    - „brafv600e“ is actually a true positive synonym for „v600e“
- Results without using a tagger.
- Compared to other applications of word embeddings
    - Many synonyms for one word
    - Very rare occuring words are used
    - Dedicated language and format



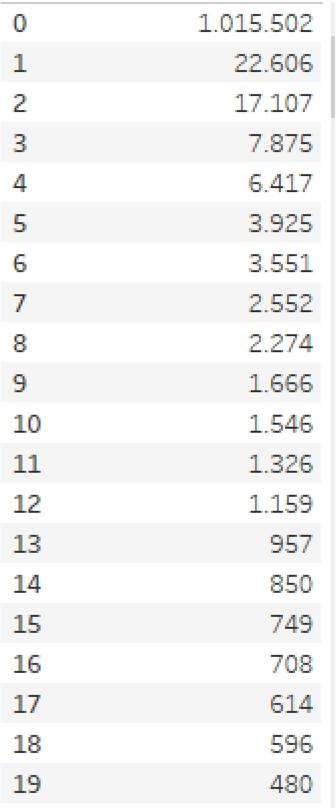
Words not appearing in PubMed&PMC

Words not appearing in model due to min-count

Log!

Error analysis

- The most mutation mentions are not even occuring in corpus
- Many mutation mentions are rarely occuring



0	1.015.502
1	22.606
2	17.107
3	7.875
4	6.417
5	3.925
6	3.551
7	2.552
8	2.274
9	1.666
10	1.546
11	1.326
12	1.159
13	957
14	850
15	749
16	708
17	614
18	596
19	480

Conclusion

- ClinVar contains more data than expected; 1,422,369 synonyms in total, that‘s 7,57% of all words in Pubmed, PMC & ScienceDirect
- The „synonym“ relationship for genetic mutations cannot be easily extracted by word embeddings
- Using a cleaned text improves the results
- Approaches where tagged entities are linked using e.g. ClinVar will outperform this method

Outlook

- Use the created word embedding models
  - on target classes with less variability (genes, diseases)
  - and try finding common dimensions that classify a token as a mutation
- Investigate further on
  - vector dimensionality
  - context size
  - better cleansing
  - more input data
  - lower min-count

Doing now what patients need next

Backup Slides

Technical hurdles – data skew

- Many short articles/abstracts

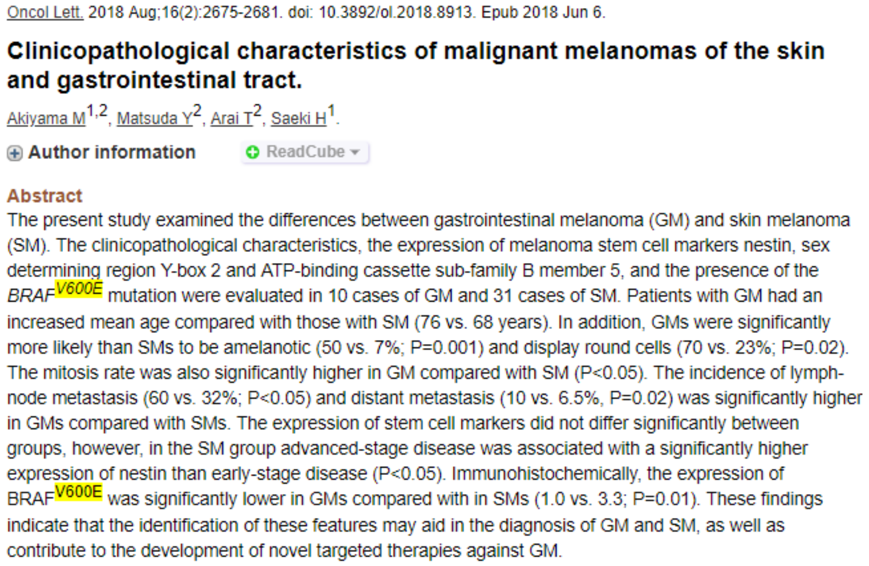
> 22 million, <1.6k characters

- Few long articles

> 1.2 million articles, >16k characters

- Rare very long articles

~ 630k articles, > 33k characters



Oncol Lett. 2018 Aug;16(2):2675-2681. doi: 10.3892/ol.2018.8913. Epub 2018 Jun 6.

Clinicopathological characteristics of malignant melanomas of the skin and gastrointestinal tract.

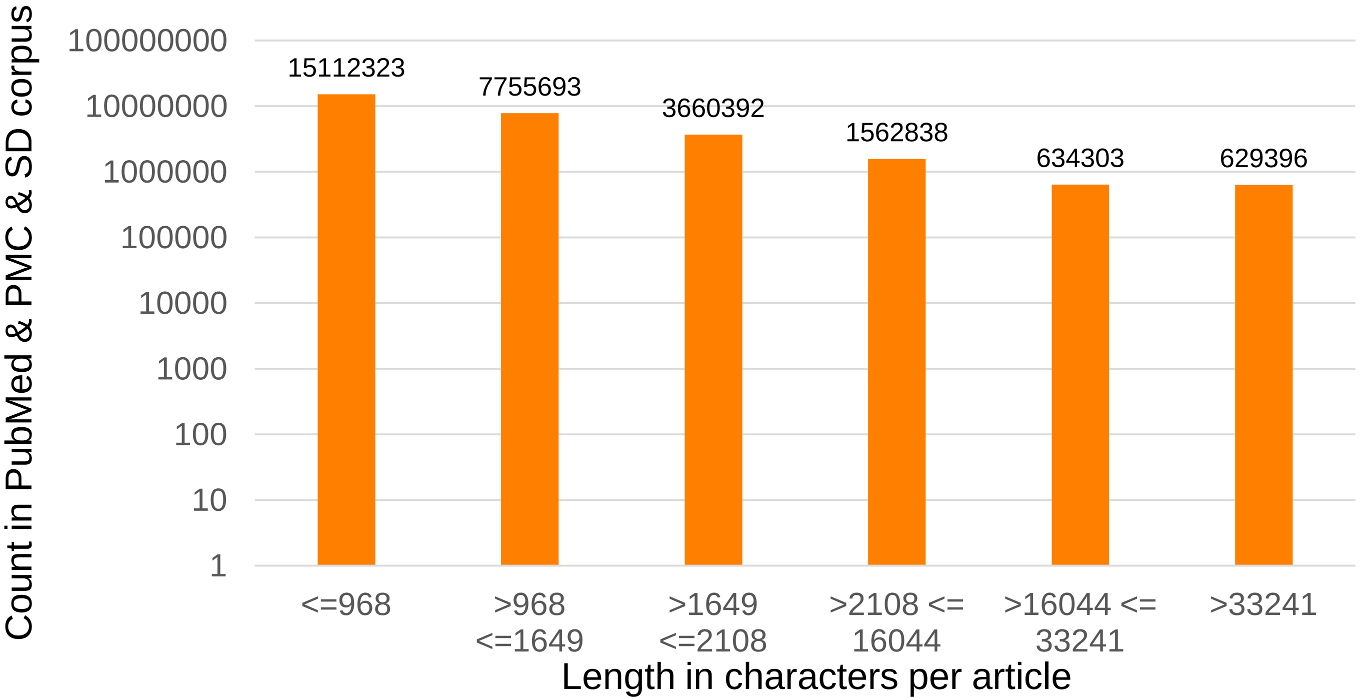
Akiyama M[1,2], Matsuda Y[2], Arai T[2], Saeki H[1].

Abstract

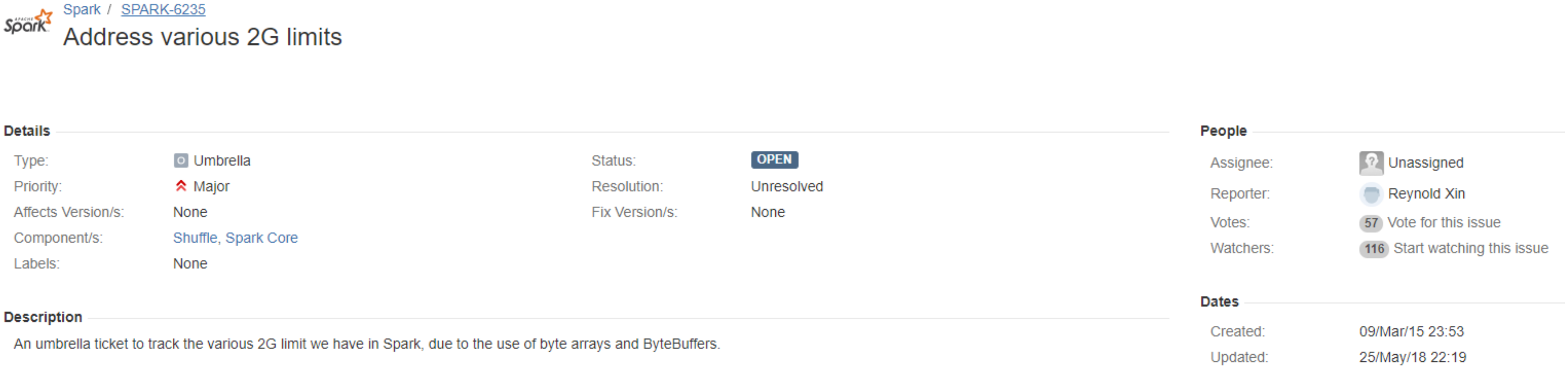
The present study examined the differences between gastrointestinal melanoma (GM) and skin melanoma (SM). The clinicopathological characteristics, the expression of melanoma stem cell markers nestin, sex determining region Y-box 2 and ATP-binding cassette sub-family B member 5, and the presence of the *BRAF*^{V600E} mutation were evaluated in 10 cases of GM and 31 cases of SM. Patients with GM had an increased mean age compared with those with SM (76 vs. 68 years). In addition, GMs were significantly more likely than SMs to be amelanotic (50 vs. 7%; P=0.001) and display round cells (70 vs. 23%; P=0.02). The mitosis rate was also significantly higher in GM compared with SM (P<0.05). The incidence of lymph-node metastasis (60 vs. 32%; P<0.05) and distant metastasis (10 vs. 6.5%, P=0.02) was significantly higher in GMs compared with SMs. The expression of stem cell markers did not differ significantly between groups, however, in the SM group advanced-stage disease was associated with a significantly higher expression of nestin than early-stage disease (P<0.05). Immunohistochemically, the expression of BRAF^{V600E} was significantly lower in GMs compared with in SMs (1.0 vs. 3.3; P=0.01). These findings indicate that the identification of these features may aid in the diagnosis of GM and SM, as well as contribute to the development of novel targeted therapies against GM.

1476 characters

Technical hurdles - data skew



Is this big data?



Spark / SPARK-6235

Address various 2G limits

Details

Type:	Umbrella	Status:	OPEN
Priority:	Major	Resolution:	Unresolved
Affects Version/s:	None	Fix Version/s:	None
Component/s:	Shuffle, Spark Core		
Labels:	None		

Description

An umbrella ticket to track the various 2G limit we have in Spark, due to the use of byte arrays and ByteBuffers.

People

Assignee:	Unassigned
Reporter:	Reynold Xin
Votes:	57 Vote for this issue
Watchers:	116 Start watching this issue

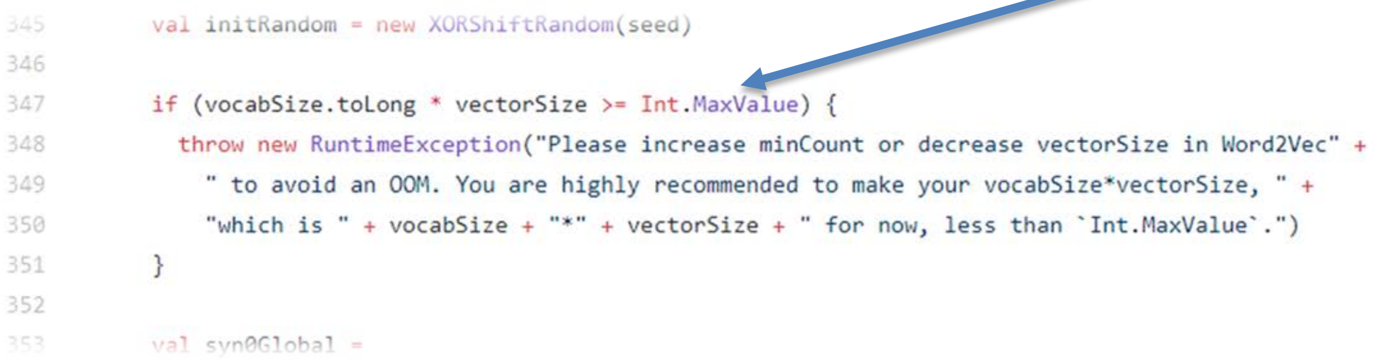
Dates

Created:	09/Mar/15 23:53
Updated:	25/May/18 22:19

Technical limitations

Spark mlib implementation of Word2Vec

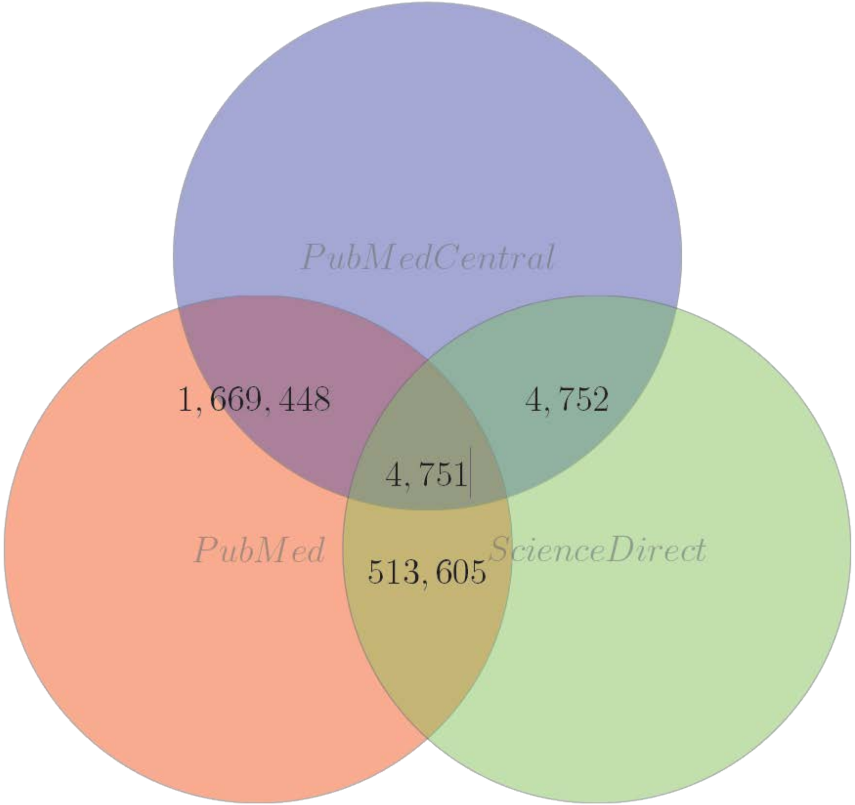
2.147.483.647 / 400 dimensions ~ 5.300.000 words



```
val initRandom = new XORShiftRandom(seed)

if (vocabSize.toLong * vectorSize >= Int.MaxValue) {
  throw new RuntimeException("Please increase minCount or decrease vectorSize in Word2Vec" +
    " to avoid an OOM. You are highly recommended to make your vocabSize*vectorSize, " +
    "which is " + vocabSize + "*" + vectorSize + " for now, less than `Int.MaxValue`.")
}

val syn0Global =
```


PubMedCentral
1,669,448
4,752
4,751
PubMed
513,605
ScienceDirect

Evaluation of the embedding model
Subsetting a „gold standard“

- ClinVar hgvs4variation / cross_references (accessed 2018-07-12 14:00)
- V600E:
  - c.1799T>A
  - p.Val600Glu
  - rs113488022
- With single letter and three-letter amino-acid codes, as well as with and without qualifier
- There are 350.832 records in in the Evaluation Set

Genetic variant extraction until today

Framework Name	MutationFinder	SETH	nala	tmVar2	VarDrugPub
Authors	Caporaso et al.	Thomas et al.	Cejuela et al.	Wei et al.	Kyubum Lee et al.
Year published	2007	2016	2017	2018	2018
Data (based on)	PubMed	PubMed, dbSNP, UniProt	PubMed	PubMed, ClinVar	PubMed
Methods	regex	Grammar matching, regex	CRF, Embeddings	regex, CRF, Dictionary lookup	Search engine, Embeddings, CNN / Random Forest
Extraction	Mutation	Mutation	Mutation	Mutation	Relations on Gene-Mutation-Disease
Normalization	None	regex+ db query	None	Regex + db query	regex

Rule-Based

Machine Learning

Rule-Based

How good is MutationFinder in recognizing Variants?

Precision : 0.89

Recall : 0.37

F-Measure : 0.53

334 documents of PubMed from the tmVar training set (hand annotated only for Variants)

→ Single-hit comparison (multiple matches ignored) aka. “Occurs in documents”

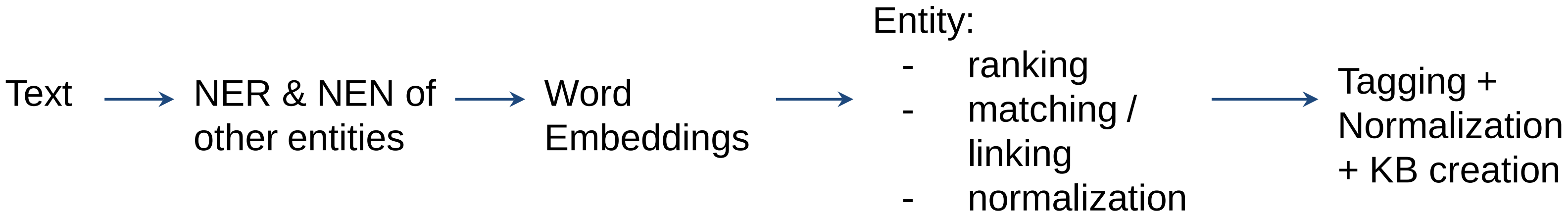
Current Situation

- Systems for extracting genes, diseases exist and are quite good
- Recent research on genetic variant extraction
  - Rule based systems
  - CRF systems
  - Normalization by regular expressions and database queries
- Data based on
  - PubMed
  - PubMedCentral (few)

What is wrong with only rule based normalization?

- ClinVar / dbSNP
    - Hand-curated
    - Manual Submission by researches

High-level Workflow



Text

NER & NEN of other entities

Word Embeddings

Entity:

- ranking
- matching / linking
- normalization

Tagging + Normalization + KB creation

Sources & Platform

- OpenAccess Data
  - Pubmed
  - PubmedCentral

- Licensed Data
  - ScienceDirect

- Hadoop / Spark
- SciBite TERMiteJ for NER
- Stanford CoreNLP for cleansing

Counts vs number of words PubMed&PMC

2	7768005
3	4663472
5	3115449
10	1946965
17	1382770
19	1289873
20	1250325

Benchmark Datasets & Tools

- Datasets
  - Fraunhofer SCAI Corpus for Normalization of Variation Mentions
  - tmVar Test https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3661051/
  - OSIRIS http://ibi.imim.es/OSIRIScorpusv01.xml
- Tools
  - Mutationfinder http://mutationfinder.sourceforge.net/
  - tmVar https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmvar/

Other research until now
genetic variant extraction

- NER/NEN for genes, diseases works great
- NER for genetic variants:
    - Precision with rule based methods is good
    - Recall with rule base methods is low
- NEN for genetic variants - Recent publications
    - NIH: tmVar2.0 *Jan 2018*
    - Lee, et al. : Deep learning of mutation-gene-drug relations from the literature *Jan 2018*
    - *Thomas et al. : SETH detects and normalizes genetic variants in text Sep 2016*

NIH : tmVar

- Enormous preprocessing (regexes)
- Conditional Random Field (CRF) for NER
- Normalization with regular expressions

But:

- No normalization from mutation to rs number
- „fine-grained rules“

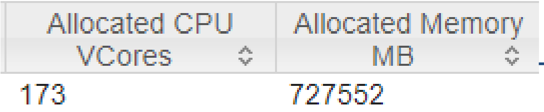
For Variants: Recall has potential Unrecognized Variants:

{('15003823', '1067-1068 ins 5 bp'), ('17671735', 'c.35delG'), ('17671735', 'p.R32W'), ('19082493', 'G/C'), ('12737948', 'IVS10+1, g-->t'), ('17671735', 'p.R127H'), ('17002658', 'g.1755 G > A'), ('17549393', 'p.Y67X'), ('18257781', 'p.F482C'), ('17549393', 'Y67X'), ('15148206', '429A-->C'), ('22125978', 'c.659_660delTA'), ('22042570', 'c.2708_2711delTTAG'), ('19370764', 'p.G204VfsX28'), ('17065190', 'C/G'), ('12737948', 'IVS3-48C'), ('14722925', 'c.87+1G>A'), ('12791036', 'R238X'), ('18257781', 'IVS21-2delAG'), ('21080147', 'E325K'), ('17169596', 'c.671G>A'), ('19592582', 'c.467C>A'), ('19110214', 'p.D2267N'), ('18257781', 'c.1445T > G'), ('12862311', '79-1 G > T'), ('20005218', 'G/A'), ('17615540', 'T87M'), ('20806042', 'p.R198W'), ('16601880', 'p.N533Y'), ('17683901', 'p.G380R'), [...] }

Multi-word-phrases

- Create a text of bigrams and count the occurences
  - Bigram construction (x2 of space)
  - Count all bigram occurences in the text
  - Count all word occurences in the text
  - Compute a score = #bigram / (#wordA + #wordB)
  - Cutoff at threshold
  - Replace Text with relevant bigrams

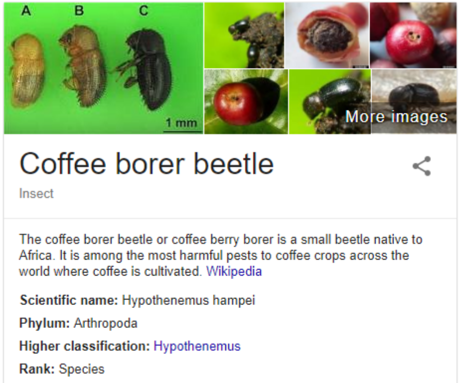
repeat



Allocated CPU VCores	Allocated Memory MB
173	727552

~ 2,2 hours for each round of construction & replacement

Phrase-Construction ⇒ Bi-grams



Coffee borer beetle

Insect

The coffee borer beetle or coffee berry borer is a small beetle native to Africa. It is among the most harmful pests to coffee crops across the world where coffee is cultivated. Wikipedia

Scientific name: Hypothenemus hampei

Phylum: Arthropoda

Higher classification: Hypothenemus

Rank: Species

2.087.023.506 Unique tokens (after NLP)

14.886.269 Unique bigrams ⇒ minOccurence 11!

20.674 bi-Grams with over 10% of co-occurence

bi-Gram generation:

- “hypothenemus hampei” more together than separated
- "latissimus dorsus” 7563 times together, 8169/10.670 individually
- “amino acid” occours 619.404, amino alone 687.695
- “significant difference” 789.247 > 10% of the cases any word is found together

Round 2

4.209 Bi-grams with over 10% of co-occurence

Examples:

1,2-bi_2-aminophenoxy ethane-n_tetraacetic

giuseppe_gasparre rodrigue_rossignol

john_wiley sons_ltd.

inferior_vena cava

spiel_ohne grenzen

Doing now what patients need next