Breast Cancer Diagnosis Text Mining Julia

Introduction

Breast cancer is the most common type of cancer diagnosed among women and is the second leading cause of cancer death. 3 Approximately 252,710 new cases of invasive breast cancer and 63,410 cases of in situ breast carcinoma were expected to be diagnosed among US women in 2017, and 40,610 women are expected to die from this disease alone.3 Overall, breast cancer mortality rates have decreased by 39%...

Load Packages and Custom Function

In [1]:
```
# init project: this will install missing packages
include("src/init.jl")

# add all packages here
using DataFrames
using Gadfly
using Vega # for word cloud
using TextAnalysis #, DimensionalityReduction, Clustering
using Plotly
using Query

# Add all custom functions here
include("src/main.jl")
using bcTextmining # our package
```

```
Plotly javascript loaded.

To load again call
init_notebook(true)


Loaded /usr/lib/jvm/default-java/jre/lib/amd64/server/libjvm.so

WARNING: Method definition get_window(Base.Dict{K, V} where V where
K) in module PlotlyJS at /home/akimaina/.julia/v0.6/PlotlyJS/src/disp
lays/electron.jl:51 overwritten at /home/akimaina/.julia/v0.6/PlotlyJ
S/src/displays/juno.jl:21.

initializing JVM and Taro...
```

Analysis

Text mining can be used to discover these knowledge patterns or hypotheses in helping to solve biomedical questions. We will perform 3 analysis i.e on treatment, diagnosis and prevention

Diagnosis

Starting by creating a dataframe of diagnosis metadata and text for the year 2008 to 2018. i.e search using medline and fetch full artcile from PMC (pubmed central) and store it as dataframe

In [3]:
```
df_full_text = @time bcTextmining.searchAndFetchFullArtcles("breast n
eoplasms","diagnosis",2008, 2018,100, true)

# # display all field except df_full_text
df_full_text[:, filter(x -> x != :fullText, names(df_full_text))]
```

```
returning cached version, to fetch afresh please set cache=false

WARNING: Compat.UTF8String is deprecated, use String instead.
  likely near In[3]:237

 19.946856 seconds (3.59 M allocations: 313.248 MiB, 4.17% gc time)
```

Out[3]:

	pmcid	pmid	date_published	title	year	pmcUrl
1	PMC2605100	19091007	20081209	Molecular imaging as a tool for translating breast cancer science.	2008	https://www.ncbi.nlm.
2	PMC2593616	19052240	20081203	Enhancing nuclear receptor-induced transcription requires nuclear motor and	2008	
3	PMC2592583	19057737	NA	Surgical images: soft tissue: An unusual presentation of perforated sigmoid	2008	
4	PMC2592581	19057734	NA	Geographic variation and physician specialization in the use of percutaneous	2008	https://www.ncbi.nlm.
5	PMC2605753	19038028	20081127	Reliable microRNA profiling in routinely processed formalin-fixed	2008	https://www.ncbi.nlm.
6	PMC2596175	19032762	20081125	Gene expression variation between distinct areas of breast cancer measured from	2008	https://www.ncbi.nlm.

	pmcid	pmid	date_published	title	year	pmcUrl
7	PMC2612673	19019216	20081119	Dose volume histogram analysis of normal structures associated with accelerated	2008	https://www.ncbi.nlm.
8	PMC2596126	19014522	20081113	Determinants of non attendance to mammography program in a region with high	2008	https://www.ncbi.nlm.
9	PMC2582941	19008355	20081113	Dynamic NMR effects in breast cancer dynamic-contrast-enhanced MRI.	2008	https://www.ncbi.nlm.
10	PMC2582583	19004780	20081112	The magnetic resonance shutter speed discriminates vascular properties of	2008	https://www.ncbi.nlm.
11	PMC2588619	19014435	20081111	Quality of life in patients with breast cancer before and after diagnosis: an	2008	https://www.ncbi.nlm.
12	PMC2588567	18990253	20081107	Mammography screening: views from women and primary care physicians in Crete.	2008	https://www.ncbi.nlm.
13	PMC2585098	18990247	20081107	The reversal of recurrence hazard rate between ER positive and negative breast	2008	https://www.ncbi.nlm.

	pmcid	pmid	date_published	title	year	pmcUrl
14	PMC2612672	18990227	20081106	Pre-segmented 2-Step IMRT with subsequent direct machine parameter optimisation –	2008	https://www.ncbi.nlm.
15	PMC2575235	18987750	20081106	Integrative Genomic Data Mining for Discovery of Potential Blood-Borne Biomarkers	2008	https://www.ncbi.nlm.
16	PMC2581822	19002271	20081104	XeNA: Capecitabine Plus Docetaxel, With or Without Trastuzumab, as Preoperative	2008	https://www.ncbi.nlm.
17	PMC2570604	18953437	NA	Overexpression of Cell Surface Cytokeratin 8 in Multidrug-Resistant MCF-7/MX	2008	https://www.ncbi.nlm.
18	PMC2570600	18953433	NA	Mammary Tumors Initiated by Constitutive Cdk2 Activation Contain an Invasive	2008	https://www.ncbi.nlm.
19	PMC2588461	18957107	20081028	Heat shock protein90 in lobular neoplasia of the breast.	2008	https://www.ncbi.nlm.
20	PMC2612006	18950515	20081025	Frequently increased epidermal growth factor receptor (EGFR) copy numbers and	2008	https://www.ncbi.nlm.

	pmcid	pmid	date_published	title	year	pmcUrl
21	PMC2577689	18947390	20081023	The clinicopathologic characteristics and prognostic significance of	2008	https://www.ncbi.nlm.
22	PMC2588622	18945363	20081022	Leptin/HER2 crosstalk in breast cancer: in vitro study and preliminary in vivo	2008	https://www.ncbi.nlm.
23	PMC2577672	18945339	20081022	Correlation of HER-2 over-expression with clinico-pathological parameters in	2008	https://www.ncbi.nlm.
24	PMC2579282	18939982	20081021	Health state utilities for non small cell lung cancer.	2008	https://www.ncbi.nlm.
25	PMC2587470	18928520	20081017	Awareness of breast cancer risk factors and practice of breast self examination	2008	https://www.ncbi.nlm.
26	PMC2575188	18925932	20081016	Tumor volume in subcutaneous mouse xenografts measured by microCT is more	2008	https://www.ncbi.nlm.
27	PMC2571108	18854030	20081014	Amplification of HER2 is a marker for global genomic instability.	2008	https://www.ncbi.nlm.
28	PMC2561063	18852895	20081014	ROCK1 and LIMK2 Interact in Spread but Not Blebbing Cancer Cells.	2008	https://www.ncbi.nlm.

	pmcid	pmid	date_published	title	year	pmcUrl
29	PMC2576333	18840272	20081007	High-resolution array CGH clarifies events occurring on 8p in carcinogenesis.	2008	https://www.ncbi.nlm.
30	PMC2567990	18837981	20081006	Identification of biomarkers in ductal carcinoma in situ of the breast with	2008	https://www.ncbi.nlm.
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Before we do any data cleaning lets make sure that the articles are relevant by plotting wordcloud

In [127]:
```
sample_size = 5 # sample 2 articels
articles_array =df_full_text[:fullText]
 wc = wordcloud(x = sample(articles_array,sample_size))
 colorscheme!(wc, palette = ("Spectral", 11))
```

Save as PNG (

Out[127]:

Data Cleaning

Looks like we have to clean and remove a few stop words.

In [4]:
```
#include("src/clean-data.jl")
# conver to corpus
arrayOfSdDoc = []
arrayOfStrText = []
for row in eachrow(df_full_text)
    sd,tx=bcTextmining.cleanText(row[:fullText])
    push!(arrayOfSdDoc,sd)
    push!(arrayOfStrText,tx)
end

# convert to corpus
corpus = Corpus(arrayOfSdDoc)
#standardize
standardize!(corpus, StringDocument)
#normalizes
#stem!(corpus) # merges words like survival and survive
```

```
WARNING: remove_nonletters! is deprecated, Use prepare! instead.
```

After cleaning...

In [129]:
```
sample_size =5 # sample size
 wc = wordcloud(x = sample(arrayOfStrText,sample_size))
 colorscheme!(wc, palette = ("Spectral", 11))
```

Save as PNG (

Out[129]:

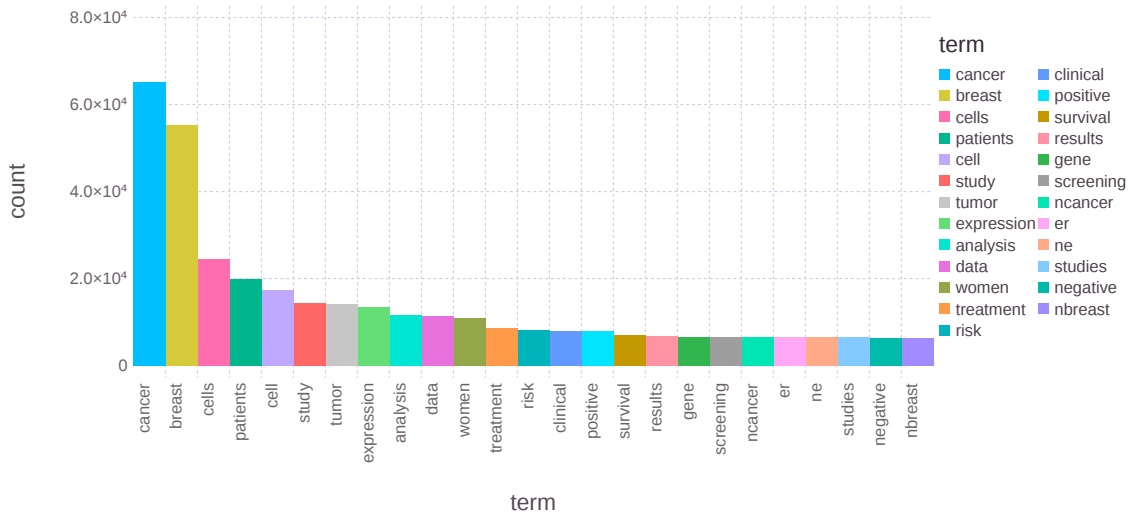
Start Analysis

In [5]:
```
set_default_plot_size(20cm, 10cm)

update_lexicon!(corpus)
#update_inverse_index!(corpus)
lexicon_df = DataFrame(term=collect(keys(corpus.lexicon)), count=coll
ect(values(corpus.lexicon)))
lexicon_df,plot =bcTextmining.fetchTopNTopic(lexicon_df,6000)
plot
```

```
  1.317531 seconds (90.67 k allocations: 4.619 MiB)
```

Out[5]:



In [131]:
```
lexicon_df

# save to csv
#writetable("output/dagnosis-lexicon.csv",lexicon_df)
```

Out[131]:

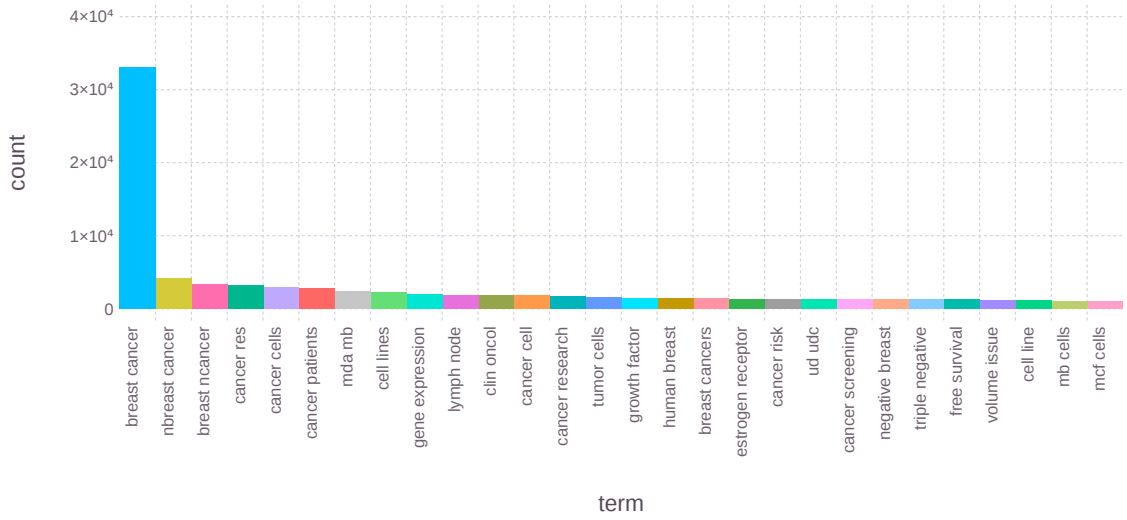
	term	count
1	cancer	65231
2	breast	55379
3	cells	24441
4	patients	19916
5	cell	17442
6	study	14391
7	tumor	14243
8	expression	13523
9	analysis	11508
10	data	11426
11	women	10949
12	treatment	8659
13	risk	8115
14	clinical	8011
15	positive	7880
16	survival	6903
17	results	6725
18	gene	6668
19	screening	6618
20	ncancer	6543
21	er	6524
22	ne	6487
23	studies	6453
24	negative	6351
25	nbreast	6230
26	receptor	5978
27	human	5917
28	time	5882
29	associated	5773
30	tissue	5770
⋮	⋮	⋮

In [6]:
```
# merge all articles
full_text_array = Array(df_full_text[:fullText])
# generate plot
df_full_n2,plot =bcTextmining.fetchNgramTopic(full_text_array,1000,2)
plot

#df_full
```

```
  0.254841 seconds (83.15 k allocations: 4.670 MiB)
```

Out[6]:



In [78]:
```
n=length(df_full_n2)
df_full_n2
# save to file
#writetable("output/dagnosis-2gram.csv",df_full_n2)
```

Out[78]: 3

mda mb cells - is a breast cancer cell

cell lines - associate with mda mb

lymph node - bc cells have been found in lymbh node

estrogen receptor - Why is knowing hormone receptor status important during daignosis https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-hormone-receptor-status.html (https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-hormone-receptor-status.html)

mcf cells -

breast carcinoma -

ductal carcinoma -

mb cells -

metastatic breast -

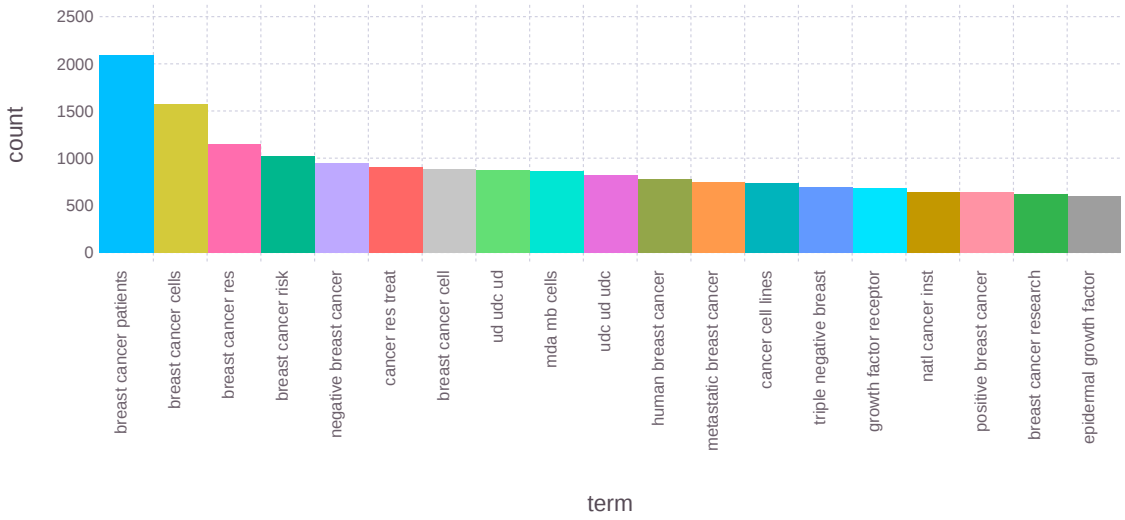
invasive breast -

triple negative - According to the status of ER, PR, HER2, breast cancer is classified as luminal A, luminal B, HER2 positive, and triple negative

In [7]:
```
# generate plot
df_full_n3,plot =bcTextmining.fetchNgramTopic(full_text_array,600,3)
plot
#df_full
```

```
  0.011178 seconds (135 allocations: 9.094 KiB)
```

Out[7]:



In [61]:
```
# save it as excel
#writetable("output/dagnosis-3gram.csv",df_full_n3)

# print head
df_full_n3
```

Out[61]:

	term	count	size
1	breast cancer patients	2091	3
2	breast cancer cells	1570	3
3	breast cancer res	1151	3
4	breast cancer risk	1020	3
5	negative breast cancer	945	3
6	cancer res treat	909	3
7	breast cancer cell	881	3
8	ud udc ud	869	3
9	mda mb cells	862	3
10	udc ud udc	823	3
11	human breast cancer	783	3
12	npage citation purposes	767	3
13	metastatic breast cancer	740	3
14	cancer cell lines	738	3
15	nhttp biomedcentral com	726	3
16	triple negative breast	694	3
17	growth factor receptor	683	3
18	natl cancer inst	646	3
19	positive breast cancer	642	3
20	breast cancer research	618	3
21	epidermal growth factor	602	3
22	clin cancer res	579	3
23	breast cancer clin	558	3
24	risk breast cancer	556	3
25	breast cancer screening	554	3
26	november volume issue	534	3
27	plosone november volume	519	3
28	patients breast cancer	506	3

HIF (Biomarker) - Hypoxia-inducible factor-1 (HIF-1) is a transcription factor that regulates gene expression in critical pathways involved in tumor growth and metastases.

induced HIF protein - HIF-1 is a crucial protein in such masses; it enables tumor progression by inducing alternative metabolic pathways within cancer cells

breast cancer cells - http://www.nationalbreastcancer.org/breast-cancer-stage-0-and-stage-1 (http://www.nationalbreastcancer.org/breast-cancer-stage-0-and-stage-1)

igf induce hif - Based on the growing body of evidence demonstrating IGF-1-induced HIF-1 activity, and thus the potential contributions of this growth

positron emission tomography - for diagnosis

reductions due mammography - for diagnosis

screening breast mri - diagnosis

nsd transgenetic females -

secondary malignant neoplasm -

hif protein accumulation -

epidermal growth factor -

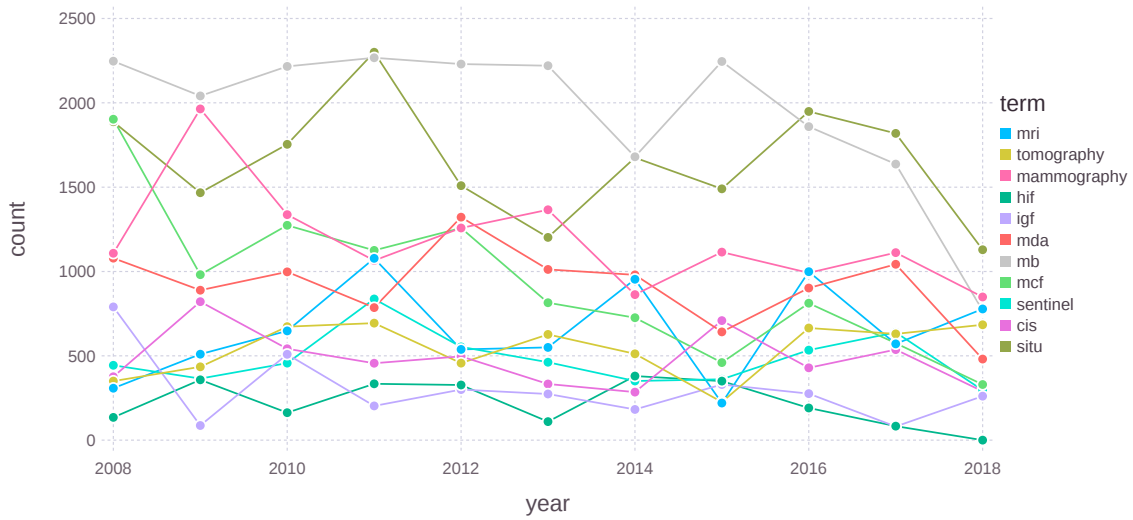
ductal carcinoma situ - Carcinoma in situ (CIS), also known as in situ neoplasm, is a group of abnormal cells. While they are a form of neoplasm, there is disagreement over whether CIS should be classified as cancer.

sentinel lymph node is defined as the first lymph node to which cancer cells are most likely to spread from a primary tumor.

Trends

In [8]:
```
# here is how to generate trends
# first define an array of terms/concepts/cells/genes e.t.c
terms = ["mri","tomography","mammography","hif","igf","mda","mb","mc
f","sentinel", "cis", "situ"]
trend_df = bcTextmining.generateTrends(df_full_text,terms,2008,2018 )
Gadfly.plot(trend_df, x="year", y="count", color="term", Geom.point,
Geom.line)
```

Out[8]:

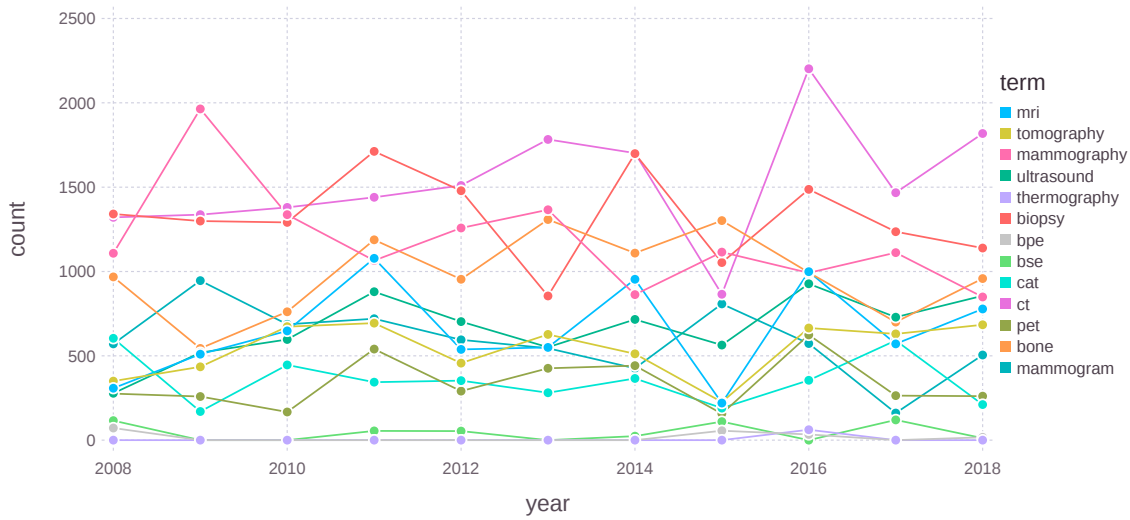


Diagnosis trends

http://www.breastcancer.org/symptoms/testing/types (http://www.breastcancer.org/symptoms/testing/types)
https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475 (https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475)

In [9]:
```
terms = ["mri","tomography","mammography", "ultrasound", "thermograph
y", "biopsy", "bpe", "bse",
        "cat", "ct", "pet", "bone", "mammogram"
        ]
trend_df = bcTextmining.generateTrends(df_full_text,terms,2008,2018 )
Gadfly.plot(trend_df, x="year", y="count", color="term", Geom.point,
Geom.line)
```

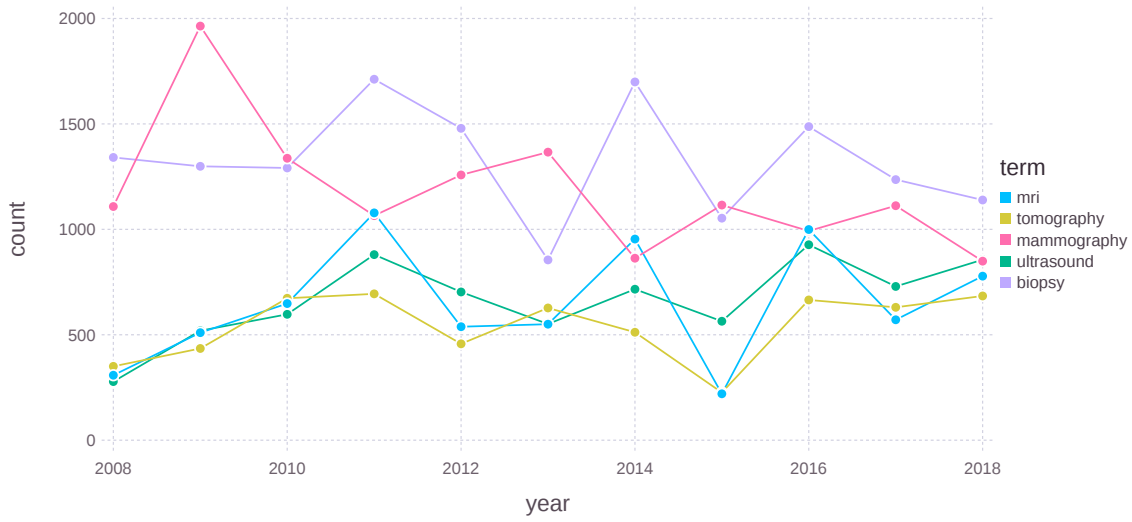
Out[9]:



Daignosis top 5 trends

In [10]:
```
# http://www.breastcancer.org/symptoms/testing/types
# https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475
terms = ["mri","tomography","mammography", "ultrasound", "biopsy"
            #, "bpe", "bse", "bone"
        ]
trend_df = bcTextmining.generateTrends(df_full_text,terms,2008,2018 )
Gadfly.plot(trend_df, x="year", y="count", color="term", Geom.point, Geom.line)
```

Out[10]:



Trends in types of Breast Cancer

http://www.nationalbreastcancer.org/breast-cancer-stage-0-and-stage-1
(http://www.nationalbreastcancer.org/breast-cancer-stage-0-and-stage-1)
https://ww5.komen.org/AboutBreastCancer/DiagnosingBreastCancer/UnderstandingaDiagnosis/TumorTypesSizes
(https://ww5.komen.org/AboutBreastCancer/DiagnosingBreastCancer/UnderstandingaDiagnosis/TumorTypesSizes

DCIS - Ductal carcinoma in situ

LCIS - Lobular carcinoma in situ

IDC - Invasive Ductal Carcinoma (IDC) - Invasive ductal carcinoma is the most common type of breast cancer (50-75 percent of all breast cancers)

Paget - disease of the nipple

ILC - InvInvasive lobular carcinoma - Invasive lobular carcinoma is the next most common type (5-15 percent of breast cancers)

metaplastic - Metaplastic breast cancer is rare, accounting for fewer than 1 percent of all invasive breast cancer

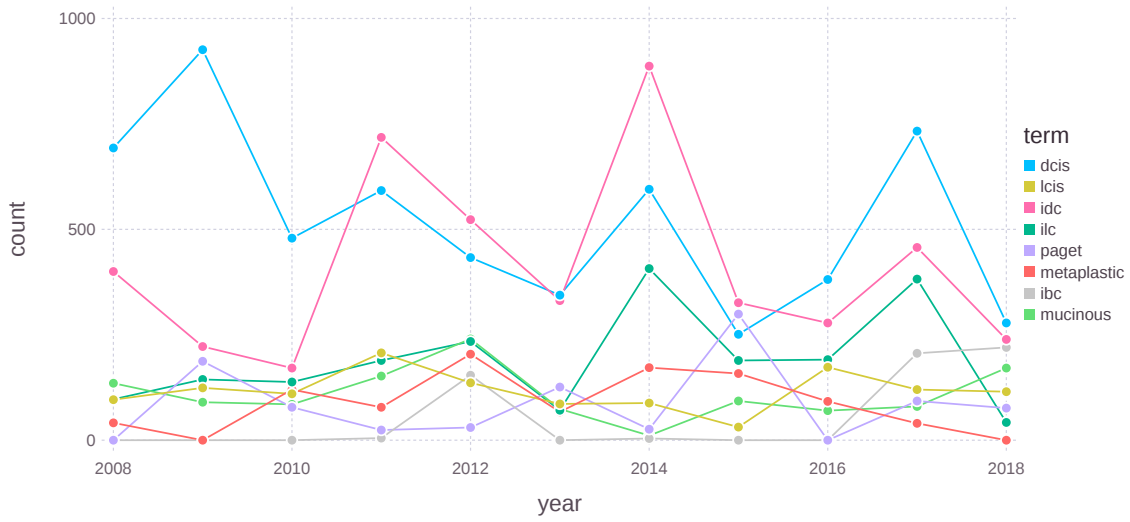
ibc - Inflammatory breast cancer (IBC) is a rare, but aggressive form of locally advanced breast cancer. main symptoms are swelling and redness

Tubular carcinoma - 1-5%

Mucinous carcinoma - 1-5%

In [11]:
```
# http://www.breastcancer.org/symptoms/testing/types
# https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475
terms = ["dcis","lcis", "idc", "ilc", "paget", "metaplastic", "ibc", "mucinous"]
trend_df = bcTextmining.generateTrends(df_full_text,terms,2008,2018 )
Gadfly.plot(trend_df, x="year", y="count", color="term", Geom.point, Geom.line)
```

Out[11]:



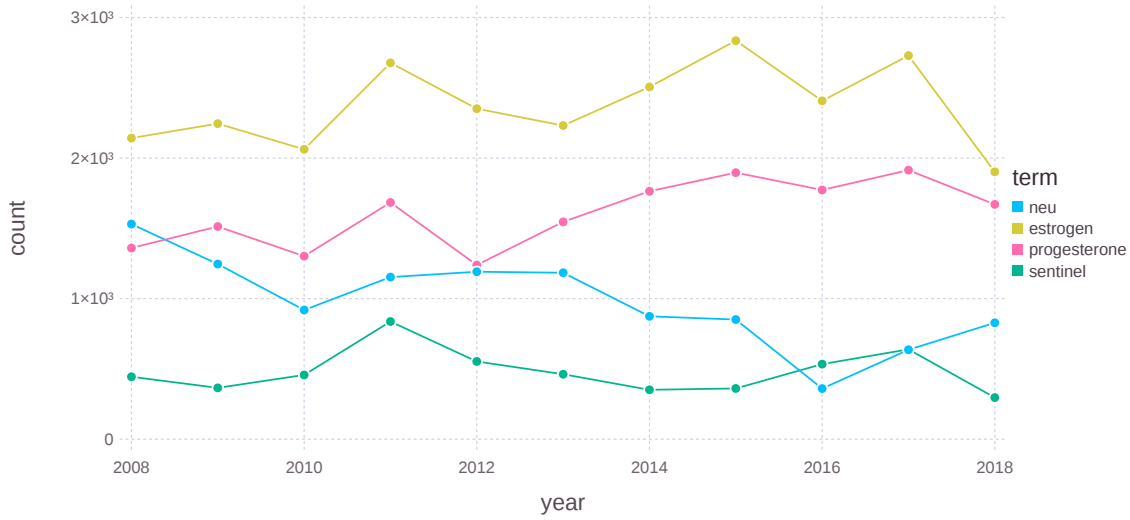
Trends in body chemicals that fuel cancer growth

http://www.nationalbreastcancer.org/growth-of-breast-cancer (http://www.nationalbreastcancer.org/growth-of-breast-cancer)

sentinel lymph node - sentinel lymph node is defined as the first lymph node to which cancer cells are most likely to spread from a primary tumor.

In [12]:
```
terms = ["neu", "estrogen", "progesterone", "sentinel"]
trend_df = bcTextmining.generateTrends(df_full_text,terms,2008,2018 )
Gadfly.plot(trend_df, x="year", y="count", color="term", Geom.point, Geom.line)
```

Out[12]:



Trends in breast cancer tumors

http://www.nationalbreastcancer.org/breast-tumors (http://www.nationalbreastcancer.org/breast-tumors)

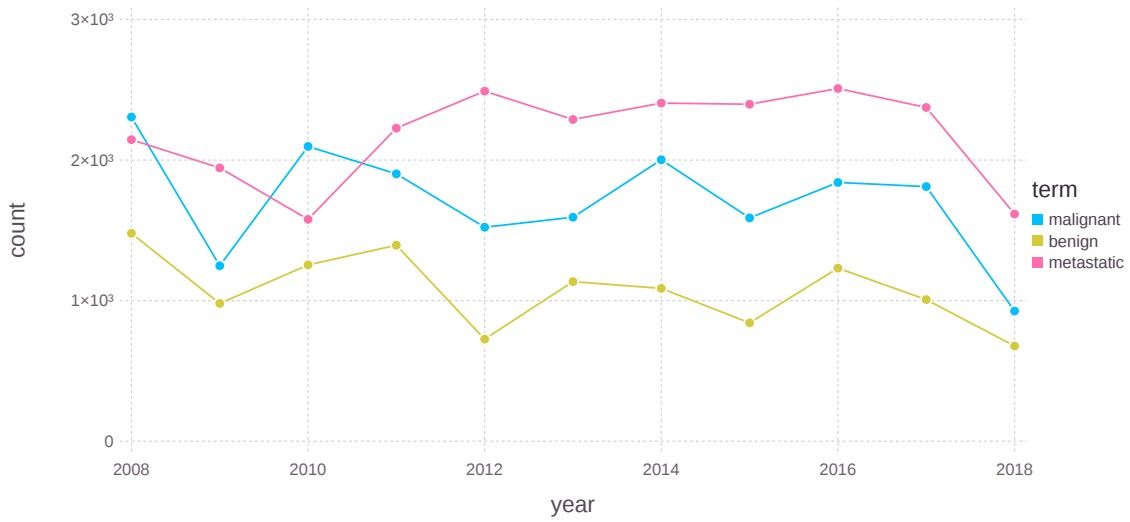
Benign Tumors - When a tumor is diagnosed as benign, doctors will usually leave it alone rather than remove it.

Malignant tumors - Malignant tumors are cancerous and aggressive because they invade and damage surrounding tissue.

Metastatic tumors - Metastatic cancer is when cancer cells of a malignant tumor spread to other parts of the body, usually through the lymph system, and form a secondary tumor..

In [13]:
```
terms = ["malignant", "benign", "metastatic"]
trend_df = bcTextmining.generateTrends(df_full_text,terms,2008,2018 )
Gadfly.plot(trend_df, x="year", y="count", color="term", Geom.point, Geom.line)
```

Out[13]:

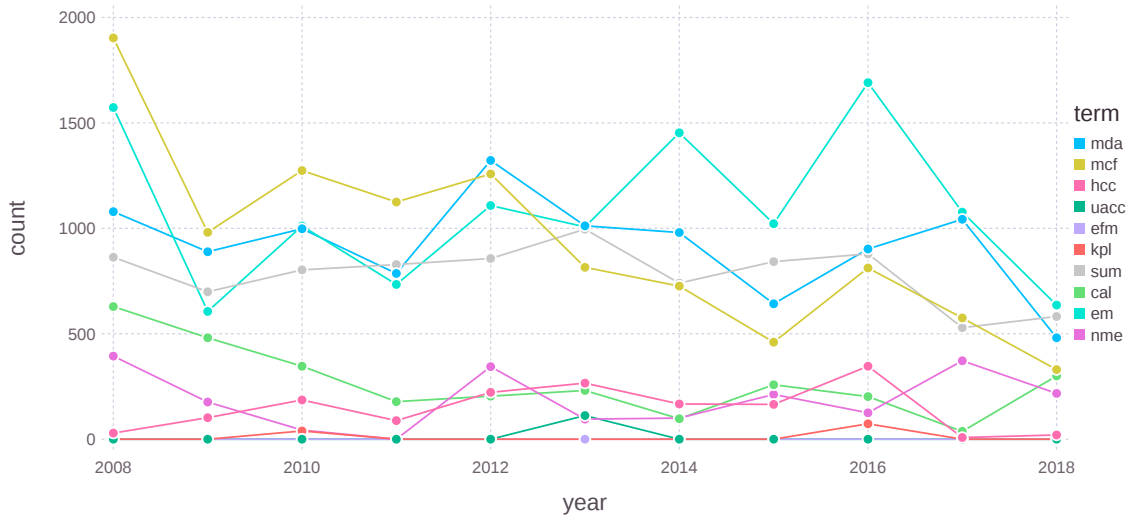


Trends in phenotypes of breast cell lines

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5665029/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5665029/) mda-mb mcf hcc

In [14]:
```
terms = ["mda", "mcf", "hcc", "uacc", "efm", "kpl", "sum", "cal", "em", "nme"]
trend_df = bcTextmining.generateTrends(df_full_text,terms,2008,2018 )
Gadfly.plot(trend_df, x="year", y="count", color="term", Geom.point, Geom.line)
```

Out[14]:



In [ ]:
```
#convert(DataFrame, d)
#plot(corpus.lexicon, x="SepalLength", y="SepalWidth", Geom.point)
#plot(x=rand(10), y=rand(10))
#NGramCorpus(corpus)

# corpus_df =convert(DataFrame, corpus)

# m = DocumentTermMatrix(corpus)

# D = dtm(m, :dense)# D
```

In [8]:
```

```

Appendix

In [9]:
```
# using R code
```