Breast Cancer Treatment 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 [2]:
```julia
# 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
```

```
checking for packages...
:) All Packages installed!
```

Plotly javascript loaded.

To load again call
```
init_notebook(true)
```

```
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.
```

```
Loaded /usr/lib/jvm/default-java/jre/lib/amd64/server/libjvm.so
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","therapy",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

 25.445806 seconds (3.59 M allocations: 313.786 MiB, 1.74% 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	https://www.ncbi.nlm.
3	PMC2592583	19057737	NA	Surgical images: soft tissue: An unusual presentation of perforated sigmoid	2008	https://www.ncbi.nlm.
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 [6]:
```
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))
```
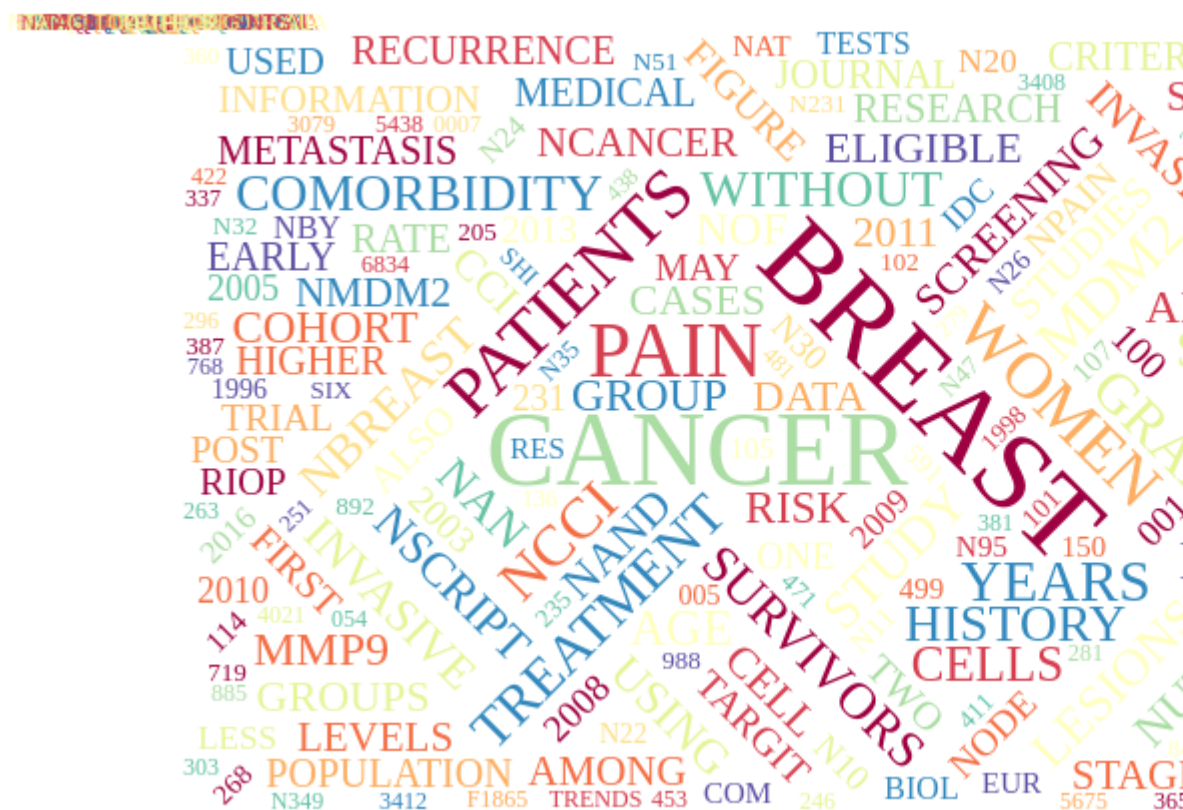

Save as PNG
(data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAA3YAAAGfCAYAAAATeUDuAAAgAF

Out[6]:

Data Cleaning

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

In [4]:
```julia
#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 [9]:
```
sample_size =5 # sample size
wc = wordcloud(x = sample(arrayOfStrText,sample_size))
colorscheme!(wc, palette = ("Spectral", 11))
```


Save as PNG
(data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAA3IAAAGfCAYAAAAakuCUAAAgAE

Out[9]:

Start Analysis

In [10]:
```
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
```

0.000155 seconds (60 allocations: 6.656 KiB)

Out[10]:
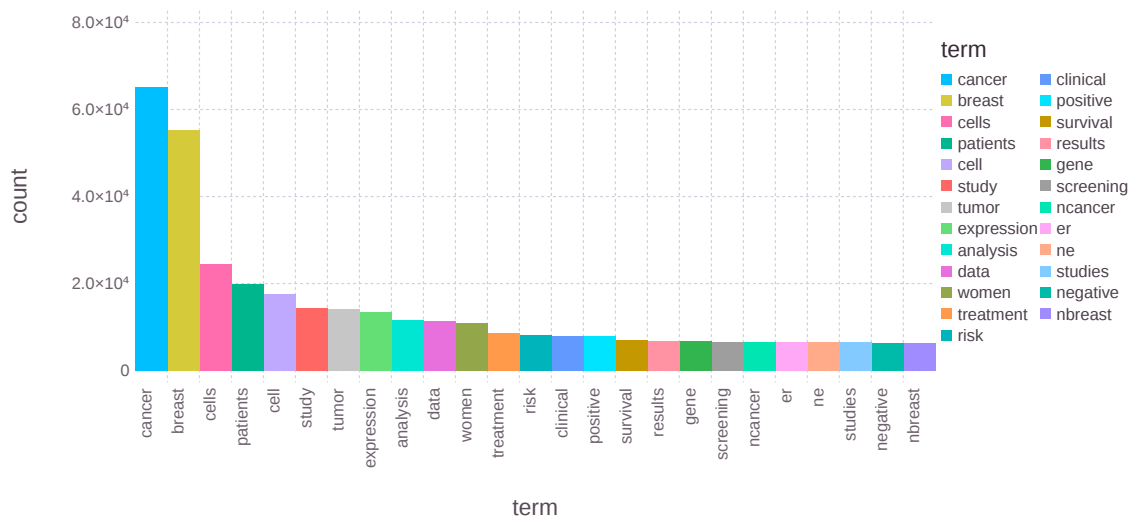
In [7]:
```
lexicon_df

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

Out[7]:

	term	count
1	cancer	71778
2	breast	61532
3	cells	26471
4	patients	22749
5	cell	18938
6	tumor	16213
7	study	15880
8	expression	15589
9	data	12760
10	analysis	12673
11	women	12435
12	treatment	9971
13	positive	9302
14	results	9194
15	clinical	9054
16	er	8962
17	risk	8701
18	figure	7919
19	survival	7684
20	screening	7593
21	negative	7507
22	gene	7444
23	studies	7202
24	receptor	6714
25	human	6639
26	associated	6504
27	time	6483
28	table	6470
29	tissue	6377
30	health	6292
⋮	⋮	⋮

In [11]:
```
# 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
```

0.310256 seconds (83.14 k allocations: 4.670 MiB)

Out[11]:
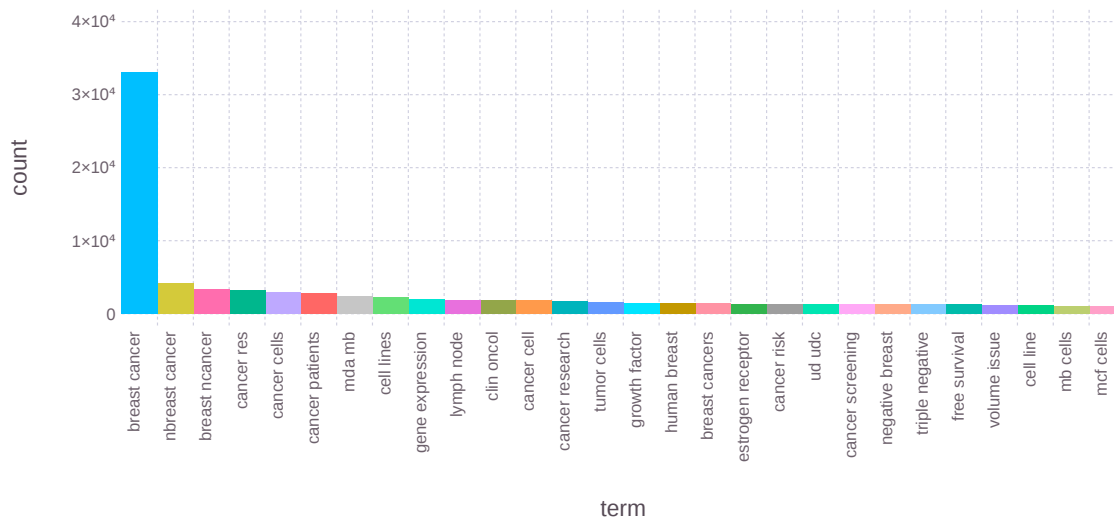
In [9]:
```
n=length(df_full_n2)
df_full_n2
# save to file
#writetable("output/dagnosis-2gram.csv",df_full_n2)
```

Out[9]:

	term	count	size
1	breast cancer	40559	2
2	ih anuscript	4416	2
3	cancer res	3906	2
4	cancer cells	3553	2
5	cancer patients	3485	2
6	anuscript ih	3015	2
7	mda mb	2988	2
8	cell lines	2635	2
9	gene expression	2556	2
10	lymph node	2366	2
11	cancer cell	2217	2
12	clin oncol	2119	2
13	tumor cells	1860	2
14	cancer research	1859	2
15	human breast	1850	2
16	growth factor	1805	2
17	estrogen receptor	1773	2
18	breast cancers	1754	2
19	cancer screening	1682	2
20	cancer risk	1588	2
21	triple negative	1536	2
22	negative breast	1534	2
23	free survival	1396	2
24	cell line	1270	2
25	bmc cancer	1259	2
26	mcf cells	1250	2
27	breast carcinoma	1249	2
28	er pr	1206	2
29	mb cells	1188	2
30	cd cd	1177	2
⋮	⋮	⋮	⋮

cd cd - The CD, or Cluster of Differential proteins are a family of type I transmembrane glycoproteins widely expressed in immune cell populations [The CD(4)(+)CD(25)(+) regulatory T cells in peripheral blood of patients with breast cancer is significantly increased in comparison with that in patients with benign breast tumor]

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 -

gene expression -

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 [12]:
```
# generate plot
df_full_n3,plot =bcTextmining.fetchNgramTopic(full_text_array,600,3)
plot
#df_full
```

0.081234 seconds (135 allocations: 9.094 KiB)

Out[12]:
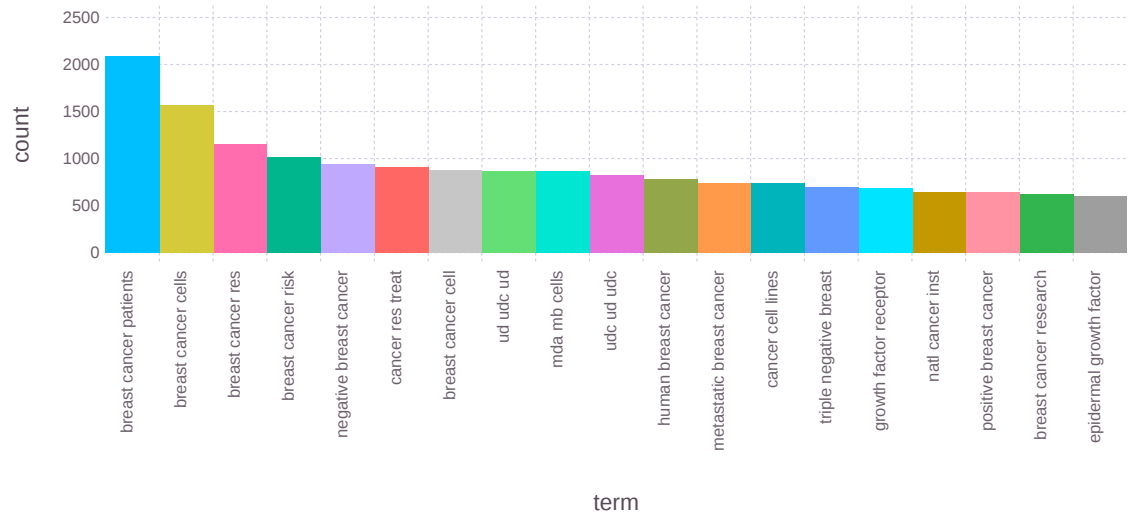
In [61]:
```python
# save it as excel
#writetable("output/dagnosis-3gram.csv",df_full_n3)

# 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

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

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.

Treatment Trends

https://www.cancer.net/cancer-types/breast-cancer/treatment-options (https://www.cancer.net/cancer-types/breast-cancer/treatment-options)

- Surgery is the removal of the tumor and some surrounding healthy tissue during an operation
- Radiation therapy is the use of high-energy x-rays or other particles to destroy cancer cells.
- Systemic therapy is treatment taken by mouth or through a vein that gets into the bloodstream to reach cancer cells wherever they may be in the body
- Chemotherapy is the use of drugs to destroy cancer cells, usually by ending the cancer cells' ability to grow and divide
- Hormonal therapy, also called endocrine therapy, is an effective treatment for most tumors that test positive for either estrogen or progesterone receptors (called ER-positive or PR-positive;
- Targeted therapy is a treatment that targets the cancer's specific genes, proteins, or the tissue environment that contributes to cancer growth and survival.

In [13]:
```
# here is how to generate trends
# first define an array of terms/concepts/cells/genes e.t.c
terms = ["surgery","radiation","chemotherapy", "hormonal", "targeted"
, "systemic"]
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]:

Surgery trends
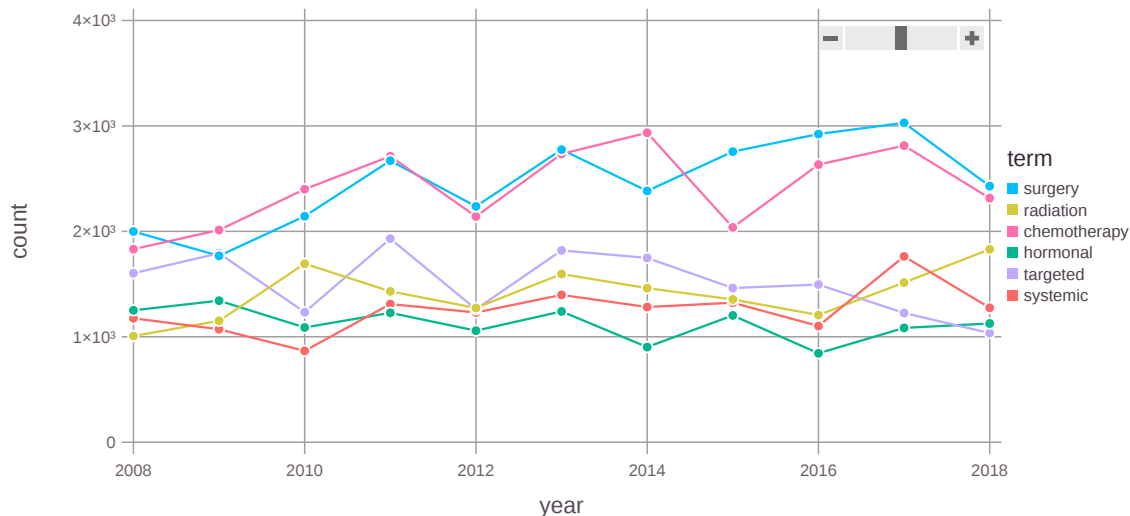
https://www.cancer.net/cancer-types/breast-cancer/treatment-options (https://www.cancer.net/cancer-types/breast-cancer/treatment-options) **Lumpectomy.** - This is the removal of the tumor and a small, cancer-free margin of healthy tissue around the tumor. Most of the breast remains. For invasive cancer, radiation therapy to the remaining breast tissue is generally recommended after surgery. For DCIS, radiation therapy after surgery may be an option depending on the patient and the tumor. A lumpectomy may also be called breast-conserving surgery, a partial mastectomy, quadrantectomy, or a segmental mastectomy.

Mastectomy. -This is the surgical removal of the entire breast. There are several types of mastectomies. Talk with your doctor about whether the skin can be preserved, called a skin-sparing mastectomy, or the nipple, called a total skin-sparing mastectomy.

In [14]:
```
terms = ["lumpectomy","mastectomy"
         ]
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]:
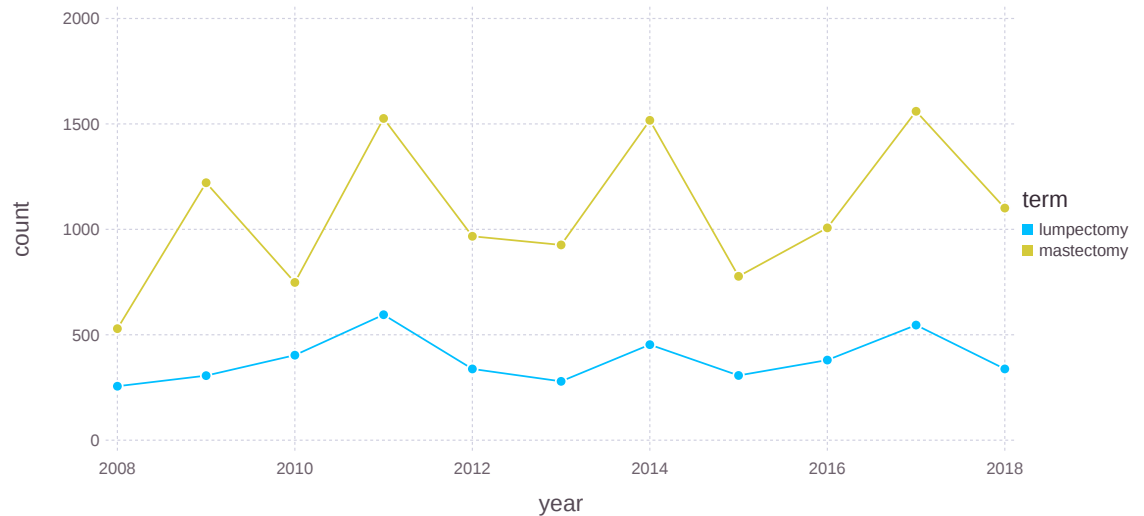
Lumpectomy surgery trends

Cancer cells can be found in the axillary lymph nodes in some cancers. It is important to find out whether any of the lymph nodes near the breast contain cancer.

Sentinel lymph node biopsy.. in a sentinel lymph node biopsy, the surgeon finds and removes a small number of lymph nodes from under the arm that receive lymph drainage from the breast. This procedure helps avoid removing multiple lymph nodes in an axillary lymph node dissection

Axillary lymph node dissection.. In an axillary lymph node dissection, the surgeon removes many lymph nodes from under the arm. These are then examined by a pathologist for cancer cells. The actual number of lymph nodes removed varies from person to person

In [15]:
```
terms = ["axillary","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[15]:
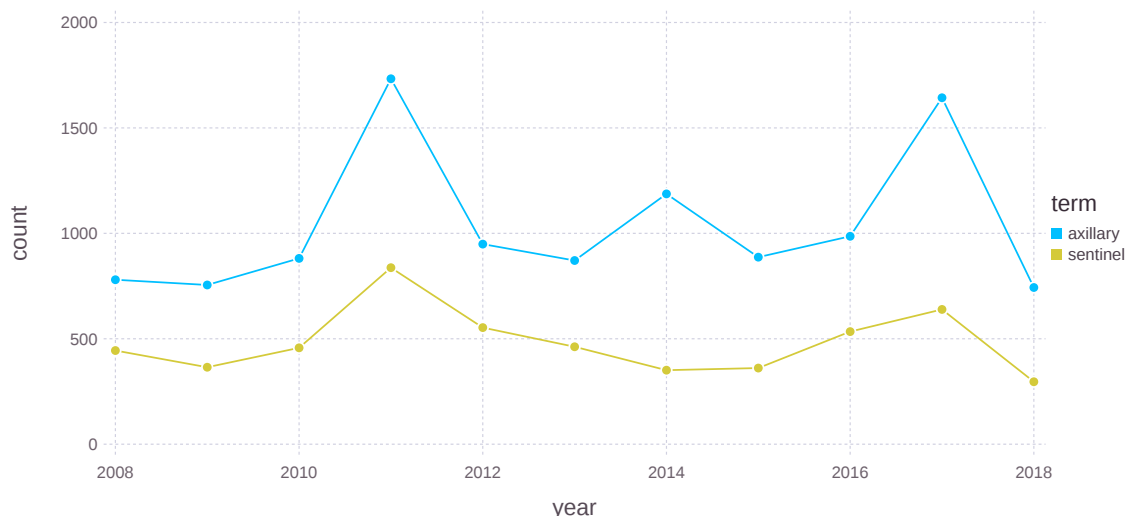
Masectomy surgery trends

Women who have a mastectomy may want to consider breast reconstruction. The techniques discussed below are typically used to shape a new breast.

Implants. A breast implant uses saline-filled or silicone gel-filled forms to reshape the breast.

Tissue flap procedures. These techniques use muscle and tissue from elsewhere in the body to reshape the breast.

Transverse rectus abdominis muscle (TRAM) flap. This method, which can be done as a pedicle flap or free flap, uses muscle and tissue from the lower stomach wall.

Latissimus dorsi flap. This pedicle flap method uses muscle and tissue from the upper back.

Deep inferior epigastric artery perforator (DIEP) flap. The DIEP free flap takes tissue from the abdomen and the surgeon attaches the blood vessels to the chest wall.

Gluteal free flap. The gluteal free flap uses tissue and muscle from the buttocks to create the breast, and the surgeon also attaches the blood vessels.

In [16]:
```
# http://www.breastcancer.org/symptoms/testing/types
# https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagno
sis-treatment/drc-20352475
terms = ["implants", "tram", "latissimus", "diep", "gluteal "
        ]
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[16]:
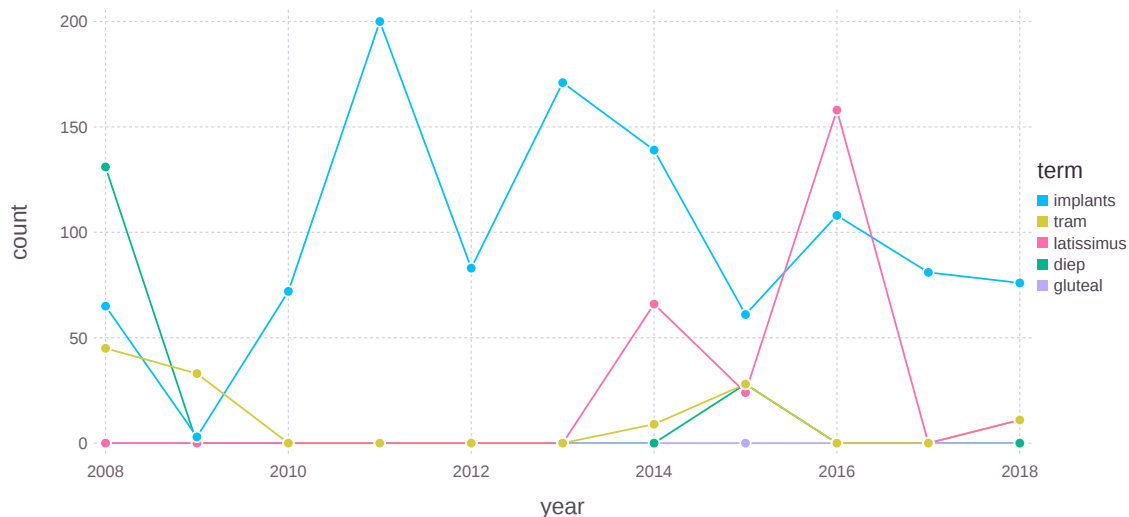
Trends in schedule of Radiation therapy

Radiation therapy may be given after or before surgery:

- Adjuvant radiation therapy is given after surgery.
- Neoadjuvant radiation therapy is radiation therapy given before surgery to shrink a large tumor,

```
In [17]:  terms = ["adjuvant","neoadjuvant"]
          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[17]:
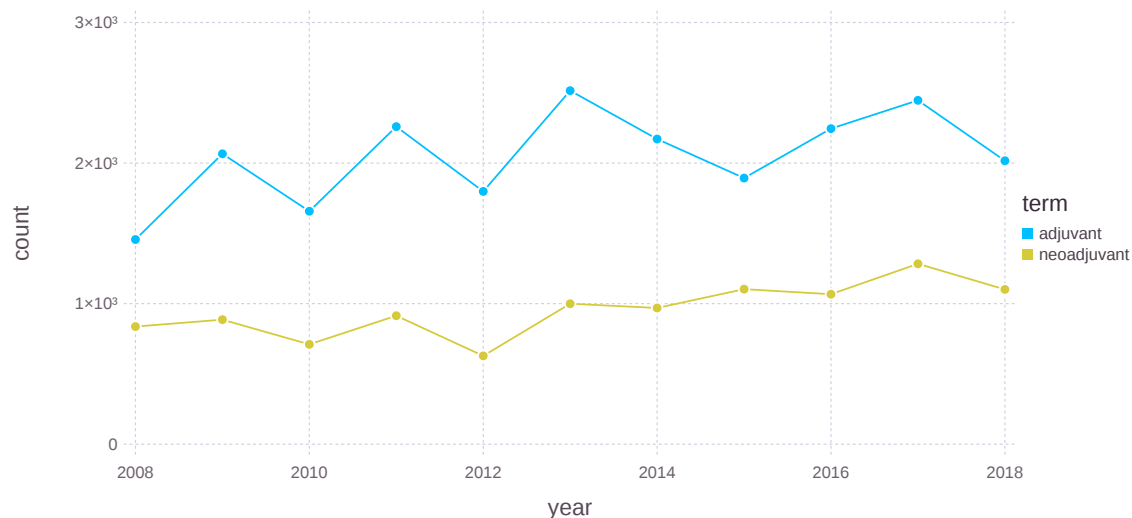
Trends in Types of Radiation therapy

irradiation (PBI) is radiation therapy that is given directly to the tumor area instead of the entire breast. It is more common after a lumpectomy..

Intensity-modulated radiation therapy (IMRT) is a more advanced way to give external-beam radiation therapy to the breast.

Proton therapy. Standard radiation therapy for breast cancer uses x-rays, also called photon therapy, to kill cancer cells. Proton therapy is a type of external-beam radiation therapy that uses protons rather than x-rays.

In [18]:
```
terms = ["irradiation", "imrt", "proton"]
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[18]:
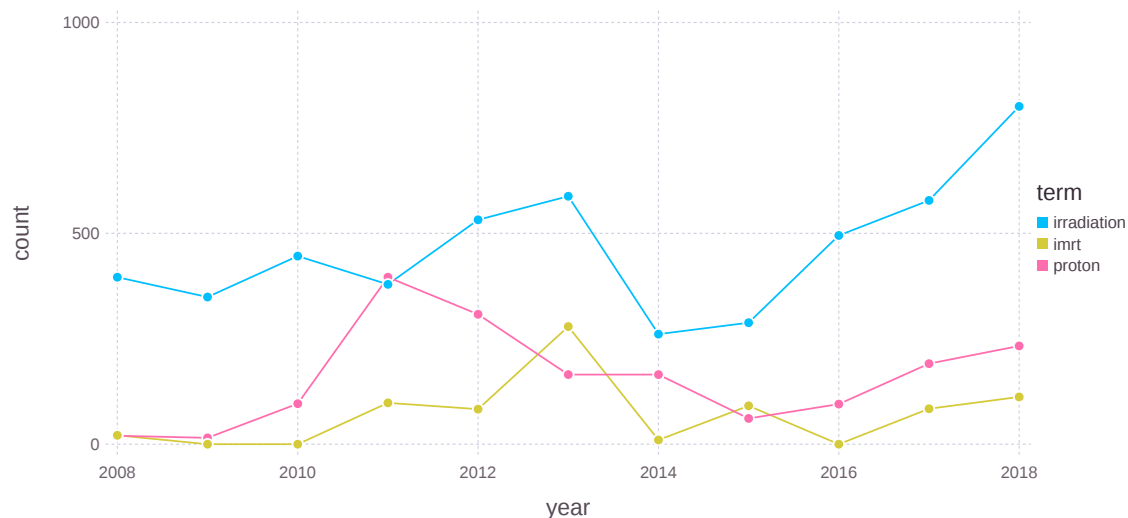
Trends in Chemotherapy

There are many types of chemotherapy used to treat breast cancer. Common drugs include: Capecitabine (Xeloda) Carboplatin (Paraplatin) Cisplatin (Platinol) Cyclophosphamide (Neosar) Docetaxel (Docefrez, Taxotere) Doxorubicin (Adriamycin) Pegylated liposomal doxorubicin (Doxil) Epirubicin (Ellence) Fluorouracil (5-FU, Adrucil) Gemcitabine (Gemzar) Methotrexate (multiple brand names) Paclitaxel (Taxol) Protein-bound paclitaxel (Abraxane) Vinorelbine (Navelbine) Eribulin (Halaven) Ixabepilone (Ixempra)

All Drugs

In [19]:
```
terms = ["capecitabine", "carboplatin", "cisplatin", "cyclophosphamid
e", "docetaxel",
        "doxorubicin", "pegylated", "epirubicin"
        , "fluorouracil", "gemcitabine", "methotrexate",
        "paclitaxel", "vinorelbine", "eribulin", "ixabepilone"
]
trend_df = bcTextmining.generateTrends(df_full_text,terms,2008,2018 )
Gadfly.plot(trend_df, x="year", y="count", color="term", Geom.point,
Geom.line)
```
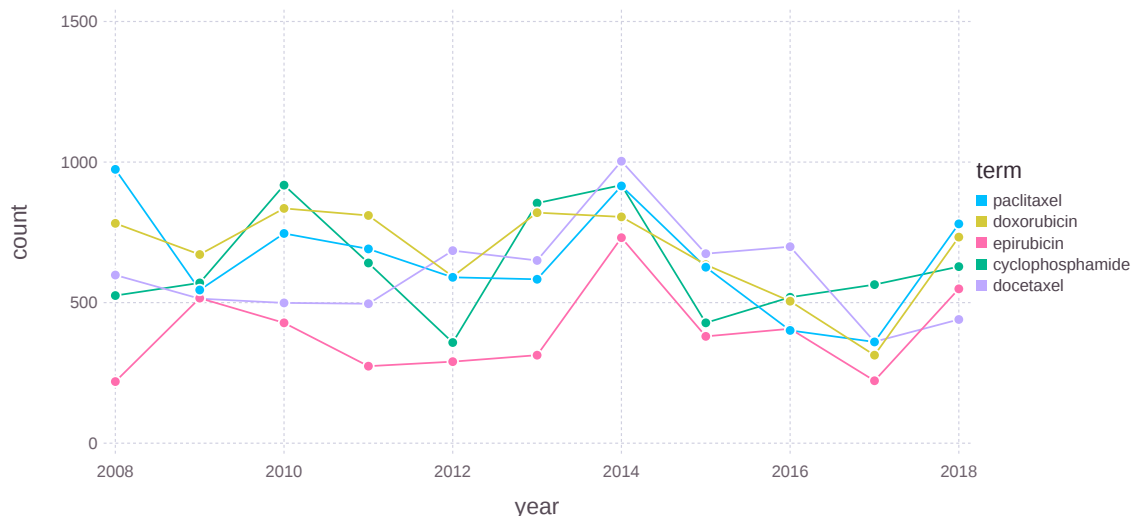
Out[19]:
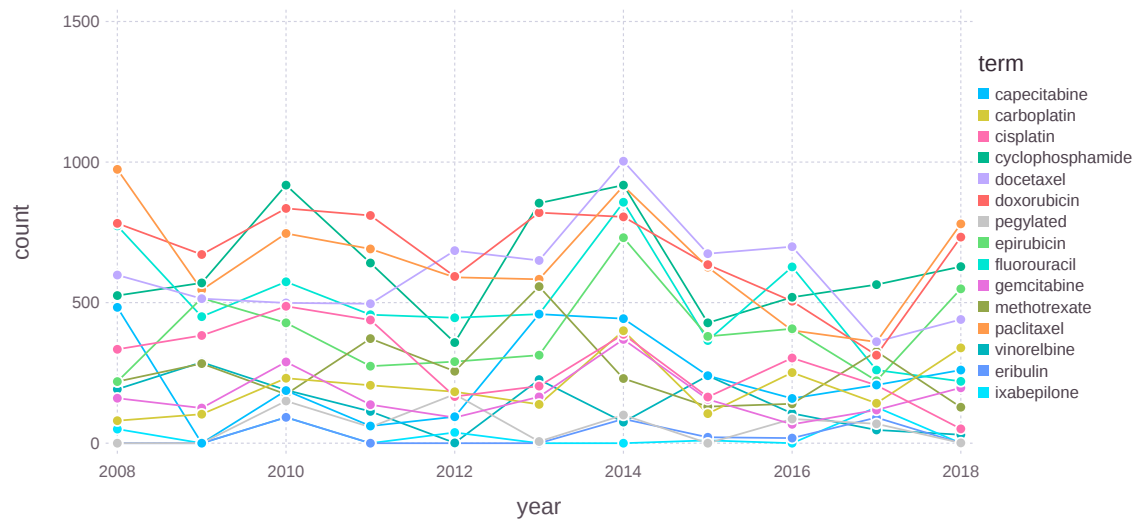
Top 5 drugs

In [20]:
```
terms = ["paclitaxel", "doxorubicin", "epirubicin", "cyclophosphamid
e", "docetaxel"
        #"doxorubicin", "pegylated", "epirubicin"
        #, "fluorouracil", "gemcitabine", "methotrexate",
        #"paclitaxel", "vinorelbine", "eribulin", "ixabepilone"
]
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[20]:

2+ Drug analysis

A patient may receive 1 drug at a time or combinations of different drugs given at the same time. Research has shown that combinations of certain drugs are sometimes more effective than single drugs for adjuvant treatment.

AC (doxorubicin and cyclophosphamide)

AC or EC (epirubicin and cyclophosphamide) followed by T (doxorubicin and cyclophosphamide, followed by paclitaxel or docetaxel, or the reverse)

CAF (cyclophosphamide, doxorubicin, and 5-FU)

CEF (cyclophosphamide, epirubicin, and 5-FU)

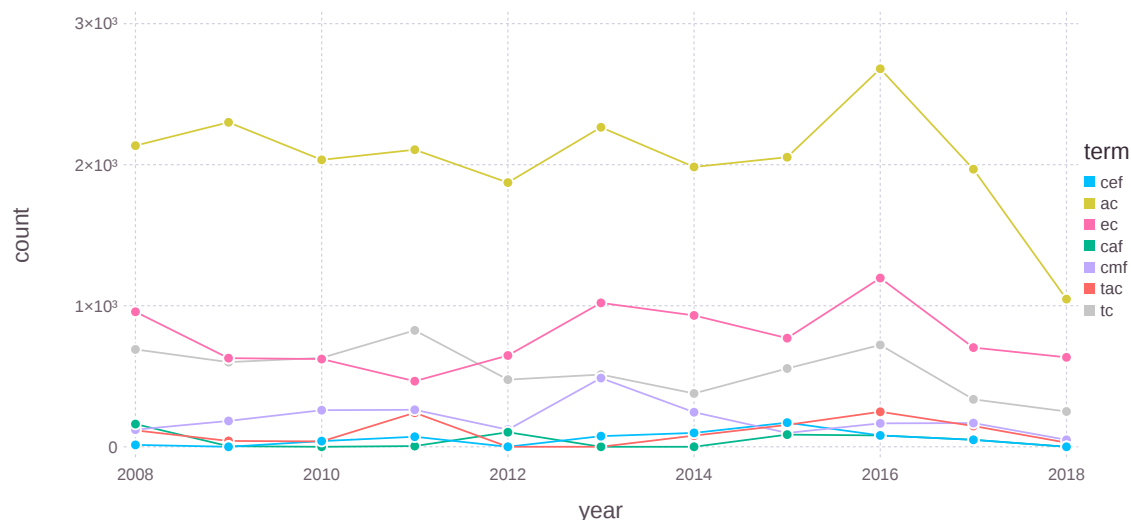CMF (cyclophosphamide, methotrexate, and 5-FU)

EC (epirubicin, cyclophosphamide)

TAC (docetaxel, doxorubicin, and cyclophosphamide)

TC (docetaxel and cyclophosphamide)

```
In [21]: terms = ["cef", "ac", "ec", "caf", "cef", "cmf", "ec", "tac", "tc"]
         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[21]:

HER2 receptor Drug analysis

Therapies that target the HER2 receptor may be given with chemotherapy for HER2-positive breast cancer (see Targeted therapy, below). An example is the antibody trastuzumab. Combination regimens for HER2-positive breast cancer may include:

ACTH (doxorubicin, cyclophosphamide, paclitaxel, trastuzumab) TCH (docetaxel, carboplatin, trastuzumab) TH (paclitaxel, trastuzumab) THP (paclitaxel or docetaxel, trastuzumab, pertuzumab) TCHP (docetaxel, carboplatin, trastuzumab, pertuzumab)

In [22]:
```
terms = ["acth", "tch", "th","thp", "tchp"]
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[22]:
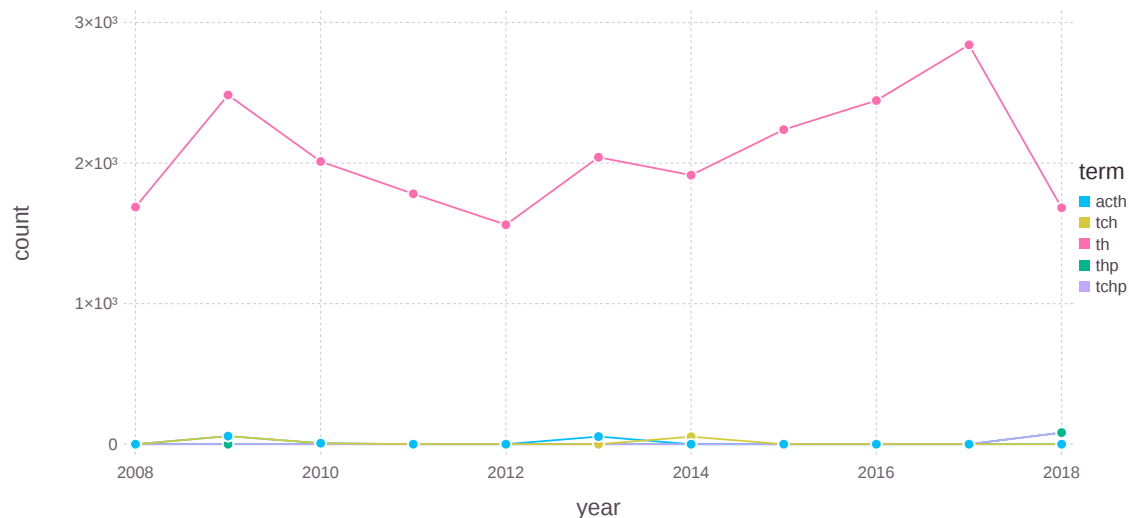
Hormonal therapy trends

Blocking the hormones can help prevent a cancer recurrence and death from breast cancer when used either by itself or after adjuvant or neoadjuvant chemotherapy.

Tamoxifen. Tamoxifen is a drug that blocks estrogen from binding to breast cancer cells. It is effective for lowering the risk of recurrence in the breast that had cancer, the risk of developing cancer in the other breast, and the risk of distant recurrence.

Aromatase inhibitors (AIs). AIs decrease the amount of estrogen made in tissues other than the ovaries in postmenopausal women by blocking the aromatase enzyme. This enzyme changes weak male hormones called androgens into estrogen when the ovaries have stopped making estrogen during menopause.

```
In [23]:  terms = ["aromatase","tamoxifen" ]
          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[23]:

Targeted therapy trends (Personalized Medicine)

Targeted therapy is a treatment that targets the cancer's specific genes, proteins, or the tissue environment that contributes to cancer growth and survival. These treatments are very focused and work differently than chemotherapy. This type of treatment blocks the growth and spread of cancer cells while limiting damage to healthy cells.

HER2-targeted therapy

Trastuzumab. This drug is approved as a therapy for non-metastatic HER2-positive breast cancer. Currently, patients with stage I to stage III breast cancer

Pertuzumab (Perjeta). This drug is approved as part of neoadjuvant treatment for breast cancer in combination with trastuzumab and chemotherapy.

Ado-trastuzumab emtansine or T-DM1 (Kadcyla). T-DM1 is a combination of trastuzumab linked to a type of chemotherapy. This allows the drug to deliver chemotherapy into the cancer cell while reducing the chemotherapy received by healthy cells. T-DM1 is approved to treat metastatic breast cancer, and studies are now testing T-DM1 as a treatment for early-stage breast cancer.

Neratinib (Nerlynx). This oral drug is approved as a treatment for higher-risk HER2-positive, early-stage breast cancer. It is taken for a year, starting after patients have finished 1 year of trastuzumab.

In [24]:
```
terms = ["trastuzumab","pertuzumab", "ado", "neratinib" ]
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[24]:
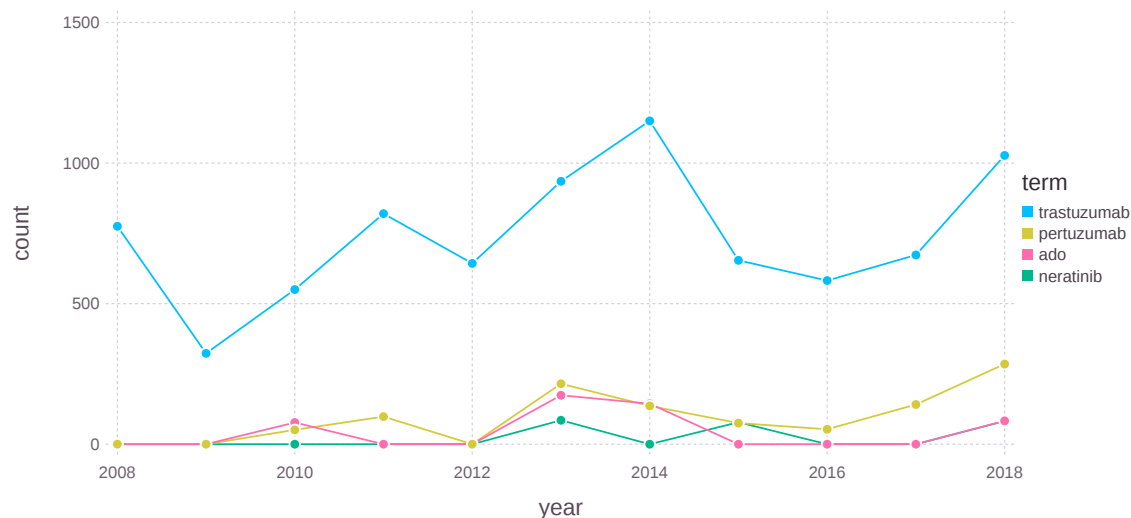
In [ ]:
```
#convert(DataFrame, d)
#plot(corpus.lexicon, x="SepalLength", y="SepalWidth", Geom.point)
#plot(x=rand(10), y=rand(10))
#NGramCorpus(corpus)

# corpus_df =convert(DataFrame, corpus)

# m = DocumentTermMatrix(corpus)

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

In [8]:

Appendix

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