


Notebook Diff

Enter notebook filenames or URLs in the form below to get started.

Please input filenames/URLs of notebooks to diff:

Base: Remote:

Hide unchanged cells 

Base

Remote

Notebook metadata changed



1	anonym	⇒⇐	1	anon
			2	

1	anonym	⇒⇐	1	anon
---	--------	----	---	------

Corpus Building, Return Migration, Historical Newspapers, Text Mining, Word Sense Disambiguation, Similarity

Metadata changed



1 Migration studies as an interdisciplinary field have become popular not only because of the increased movement of people across the world for various reasons but also because globalization has fostered and facilitated the movement of humans, goods, technology, and information. Despite its inherent transnational nature, national, but especially regional approaches, have proven to be fruitful for the study of historical developments of migration where archival material is scarce and sources often are lacking. Return migration, also termed remigration or repatriation, is still an all too frequently neglected topic within migration studies and migration history. At the same time, repatriation has always been part of every migration movement also in recent times. Since remigration studies are, in a sense, somewhat hidden within migration discourses and research we encounter even more challenging issues concerning primary sources to analyze. Therefore, historians oftentimes turn to (digitized) historical newspapers. The use of these sources comes in tandem with complex challenges, and this is where the present paper is to be positioned.

1 Migration studies as an interdisciplinary field have become popular not only because of the increased movement of people across the world for various reasons but also because globalization has fostered and facilitated the movement of humans, goods, technology, and information. Despite its inherent transnational nature, national, but especially regional approaches, have proven to be fruitful for the study of historical developments of migration where archival material is scarce and sources often are lacking. Return migration – which can generally be defined as 'cross-border migration to the country of origin' <cite data-cite="6142573/YKNUBLK6"></cite> –, is still an all too frequently neglected topic within migration studies and migration history. At the same time, repatriation has always been part of every migration movement also in recent times. Since remigration studies are, in a sense, somewhat hidden within migration discourses and research, we encounter even more challenging issues concerning primary sources to analyze. Therefore, historians often turn to (digitized) historical newspapers. The use of these, however, comes in tandem with complex challenges and a necessary update on source criticism, which has gotten much attention and is discussed within the emerging field of digital hermeneutics <cite data-cite="6142573/JMZAZWUX"></cite> <cite data-cite="8918850/AH3TIH3N"></cite><cite data-cite="6142573/LAF2DBJT"></cite> <cite data-cite="6142573/5FI5SV3F"></cite>. Other parts of the historical method, like a detailed critical assessment of adequate corpus creation in the heuristic research step, have so far not been written about enough. The challenges within the heuristic research step, however, are no less. For example, ambiguous keywords can complicate the search and lead to results that are not relevant for the research question. Also, specific topics, discourses or ideas are difficult to track down by keyword searches alone.

1 Historiography based on digital sources usually comes hand in hand with the need to search for complex patterns in masses of information rather than gaps in the historical record <cite data-cite="6142573/B353HSFG"></cite>. This development has led to a certain hybridity of classical and digital methods in historical research <cite data-cite="6142573/NY822LF2"></cite>, and goes along with methodological and epistemological challenges for the historical disciplines. It also calls for an extension of the historical method, which guides historical understanding through the three steps heuristics, source criticism and interpretation <cite data-cite="8918850/AH3TIH3N"></cite>. Especially the update on source criticism has received much attention and is discussed within the emerging field of digital hermeneutics <cite data-cite="6142573/JMZAZWUX"></cite> <cite data-cite="8918850/AH3TIH3N"></cite><cite data-cite="6142573/LAF2DBJT"></cite><cite data-cite="6142573/5FI5SV3F"></cite>. Other parts of the historical method, like a detailed critical assessment of adequate corpus creation in the heuristic research step, have so far not been written about enough.

1 The motivation of this paper is to present and describe a digital workflow that goes from building and refining a newspaper corpus using text mining methods to the qualitative analysis of the final results. In particular, the paper shows how a corpus created with ambiguous search queries was successfully classified into relevant and irrelevant articles, i.e., disambiguated by applying digital methods, and how the final corpus was used for a further qualitative, discourse-driven analysis of return migration from the Americas to Europe between 1850 and 1950. In doing so, our overall goal is to underline the necessity to give more thought and research to support digital methods that lie between qualitative analysis of small information units and quantitative approaches to big data – we call it the meso level. The search for complex patterns in masses of information rather than gaps in the historical record <cite data-cite="6142573/B353HSFG"></cite> has led to a hybridity of classical and digital methods in historical research <cite data-cite="6142573/NY822LF2"></cite>. The aim of the mesoanalysis proposed here is to sort text extracted from a large corpus of data and, e.g., sort it according to topics, content or actors using automated methods thus creating a research driven corpus for further analysis, such as discourse analysis. Discourse analysis has always also relied on a thorough reading of relevant text corpora. With ever-increasing large digital datasets, automated corpus-specific approaches (such as the calculation of multi-words units) support qualitative interpretative steps like the ones needed for discourse analysis <cite data-cite="6142573/KAFFLBWQ"></cite><cite data-cite="6142573/GRIVXPM6"></cite>. All in all, we present a corpus building method that supports humanities research, which means that the focus lies on solving a specific problem and not on a comparison or evaluation of different approaches or methods. Still, the methodology presented in this paper can be adapted for research projects that deal with similar corpus building issues.

1 Historical research is often driven by event- or topic-specific research questions. For research on return migration – which can generally be defined as 'cross-border migration to the country of origin' <cite data-cite="6142573/YKNUBLK6"></cite> –, we started with questions on how Austrian newspapers reported on return migration to Europe between 1850 and 1950, what kind of discourses can be found and how they developed over time. This means that although using big data and quantitative methods to find patterns that overlap with these research questions can be rewarding, especially for discourse related issues it oftentimes is still necessary to find and extract those parts in the massive data dumps that are relevant for the topic in question. For this reason, corpus building is an essential aspect of working with large amounts of digital sources. However, creating good corpora often requires time-consuming and complex search processes. In order to find articles on the topic of return migration, it is first necessary to find keywords that actually return articles on the topic. Then, it must be checked (by close reading) whether the keyword search omits important articles. If this is so, more and also broader search terms have to be included, which in turn can lead to articles being found that are not relevant to the research question posed. For example, the German term 'Rückwanderer' (*returnee*) returns only relevant articles but not all relevant articles available. The German term for return migration ('Rückwanderung'), on the other hand, has different meanings in different contexts and returns too many irrelevant texts. This is very daunting because it makes it necessary to weigh up between a collection that misses relevant articles, and one which contains noise (i.e. irrelevant texts) <cite data-cite="6142573/RZWYTHC7"></cite><cite data-cite="6142573/TTCX55K3"></cite>.

⇒⇐

1 Like our project on return migration, historical research is often driven by event- or topic-specific research questions. We started with questions on how Austrian newspapers reported on return migration to Europe between 1850 and 1950, what kind of discourses can be found and how they developed over time. This means that although using big data and quantitative methods to find patterns that overlap with these research questions can be rewarding, especially for discourse related issues it oftentimes is still necessary to find and extract those parts in the massive data dumps that are relevant for the topic in question. For this reason, corpus building is an essential aspect of working with large amounts of digital sources. However, creating good corpora often requires time-consuming and complex search processes. In order to find articles on the topic of return migration, it is first necessary to find keywords that actually return articles on the topic. Then, it must be checked (by close reading) whether the keyword search omits important articles. If this is so, more and also broader search terms have to be included, which in turn can lead to articles being found that are not relevant to the research question posed. For example, the German term 'Rückwanderer' (returnee) returns only relevant articles but not all relevant articles available. The German term for return migration ('Rückwanderung'), on the other hand, has different meanings in different contexts and returns too many irrelevant texts. This is very daunting because it makes it necessary to weigh up between a collection that misses relevant articles, and one which contains noise (i.e. irrelevant texts)<cite data-cite="6142573/RZWYTHC7"></cite><cite data-cite="6142573/TTCX55K3"></cite>.

1 This paper shows how a corpus created with ambiguous search queries related to return migration is successfully classified into relevant and irrelevant articles, i.e., disambiguated by applying digital methods. A semi-supervised similarity-based word sense disambiguation (WSD) approach using Latent Dirichlet allocation (LDA), a probabilistic model that calculates the probability distribution over terms <cite data-cite="6142573/CVSNSE2"></cite>, and the Jensen-Shannon (JS) distance (the square of the Jensen-Shannon divergence), which measures the similarity between texts <cite data-cite="6142573/LM8L24CE"></cite>, was applied to reach this goal. The ability to deal with complex, large-scale collections with different themes and without a clear boundary between relevant and non-relevant texts has made it our preferred method.

2 Both approaches, the training of the LDA algorithm as well as the similarity measurements are unsupervised and build on the whole context of a document. LDA topics can capture the polysemous or ambiguous use of words, but they do not carry the explicit notion of the correct context that is necessary for WSD <cite data-cite="6142573/WLBU3DX"></cite>. Therefore, a training/feedback corpus with information (labels) on the 'correct' or 'incorrect' context (relevant or irrelevant for the research question) was created manually for document comparison and clustering. The document labels do not play a role in training the LDA algorithm and finding the most similar set of documents in the feedback corpus, however, they allow the calculation of the overall relevance of the retrieved most similar set of documents based on the numeric labels. This calculation is used to support the final classification into relevant and non-relevant documents, as explained step by step in the hermeneutics layers of this paper.

1 We chose a semi-supervised similarity-based word sense disambiguation (WSD) approach using Latent Dirichlet allocation (LDA), a probabilistic model that calculates the probability distribution over terms <cite data-cite="6142573/CVSNSE2"></cite>, and the Jensen-Shannon (JS) distance (the square of the Jensen-Shannon divergence), which measures the similarity between texts <cite data-cite="6142573/LM8L24CE"></cite>, was applied to reach this goal. The ability to deal with complex, large-scale collections with different themes and without a clear boundary between relevant and non-relevant texts has made it our preferred method.

2 Both approaches, the training of the LDA algorithm as well as the similarity measurements are unsupervised and build on the whole context of a document. LDA topics can capture the polysemous or ambiguous use of words, but they do not carry the explicit notion of the correct context that is necessary for WSD <cite data-cite="6142573/WLBU3DX"></cite>. Therefore, a training/feedback corpus with information (labels) on the 'correct' or 'incorrect' context (relevant or irrelevant for the research question) was created manually for document comparison and clustering. The document labels do not play a role in training the LDA algorithm and finding the most similar set of documents in the feedback corpus, however, they allow the calculation of the overall relevance of the retrieved most similar set of documents based on the numeric labels. This calculation is used to support the final classification into relevant and non-relevant documents, as explained step by step in the hermeneutics layers of this paper.

1 The motivation of this paper is to present and describe this process by means of a specific topic (migration) and a specific research question (concerning return migration). In addition, we want to show how the created corpus was used to conduct a qualitative, discourse-driven historical analysis on return migration from the Americas to Europe between 1850 and 1950. We want to underline the necessity to investigate more thought and research to support digital methods that lie between qualitative analysis of small information units and quantitative approaches to big data – we call it the meso level. The aim of the mesoanalysis proposed here is to sort text extracted from a large corpus of data and, e.g., sort it according to topics, content or actors using automated methods thus creating a research driven corpus for further analysis.

1 Discourse analysis has always also relied on a thorough reading of relevant text corpora. With ever-increasing large digital datasets, automated corpus-specific approaches (such as the calculation of multi-words units) support qualitative interpretative steps like the ones needed for discourse analysis <cite data-cite="6142573/KAFFLBWQ"></cite><cite data-cite="6142573/GRIVXPM6"></cite>. In doing so, we present a corpus building method that supports humanities research, which means that the focus lies on solving this one specific problem and not on a comparison or evaluation of different approaches or methods. Still, the methodology presented in this paper can be adapted for research projects that deal with similar corpus building issues.

In [1]:

In [2]:

```
(...)  
6     "type": "image",  
7     "source": [  
8         "figure 1: Das  
interessante Blatt, 12.12.1907, p.  
1. ",  
9     ]  
10    }  
(...)
```

```
(...)  
6     "type": "image",  
7     "source": [  
8         "figure 1: Das  
interessante Blatt, 12.12.1907, p.  
1. ",  
9     ]  
10    }  
(...)
```

Outputs unchanged



In [2]:

In [3]:

```

1 metadata_2={
2     "jdh":{
3         "module": "object",
4         "object": {
5             "type": "image",
6             "source": [
7                 "figure 2: Salzburger Blatt, 5.5.1946, p. 8.",
8             ],
9         }
10    }
11 }
12 }
13 display(Image("images/rückwanderer2.png"), metadata=metadata_2)

```

Outputs unchanged

In [3]:

In [4]:

```

1 import pandas as pd
2 import re
3 import re, numpy as np, pandas as pd
4 import csv
5 from pprint import pprint
6 from IPython.display import display
7 get_ipython().magic(u'matplotlib inline')
8 #import data
9 df = pd.read_csv('data/export_returnmigration_16_04_2021_21_35.csv', usecols =
10 ['text', 'relevancy'])
11 display(df[22:24].style.set_caption(caption_content).hide_index(), metadata={"jdh":
12 {"object":{"source": [caption_content]}}})

```

Outputs changed

table 3: Text with relevancy labels (3 = relevant; 0 = irrelevant).

relevancy	text
0	(Ein österreichischer Asienforscher.) Der Wien=Floridsdorfer Turnlehrer Anton Gebauer befand sich seit vergangenen Herbst auf einer Forschungsreise, deren Ziel Hinterindien mit den bisher von Weißen noch nicht betretenen Schanstaaten war. Das, was bisher den größten Forschern nicht gelungen war, scheint Gebauer durch Glück, Mut und Todesverachtung möglich gewesen zu sein, denn nach seinen eingetroffenen Briefen hat er tatsächlich diese Gebiete bereist und weilte dort längere Zeit. Nun kehrte er wieder in seine Vaterstadt Bennisch in Schlesien zurück und diese Rücksicht gab Anlaß zu einer großen Ehrung für den Heimgekehrten, der nunmehr mit den größten Forschern in gleiche Reihe gestellt werden kann. Der junge Forscher wurde von der Stadtvertretung, den Vereinen und der Bevölkerung festlich empfangen.

table 3: Text with relevancy labels (3 = relevant; 0 = irrelevant).

relevancy	text
0	(Ein österreichischer Asienforscher.) Der Wien=Floridsdorfer Turnlehrer Anton Gebauer befand sich seit vergangenen Herbst auf einer Forschungsreise, deren Ziel Hinterindien mit den bisher von Weißen noch nicht betretenen Schanstaaten war. Das, was bisher den größten Forschern nicht gelungen war, scheint Gebauer durch Glück, Mut und Todesverachtung möglich gewesen zu sein, denn nach seinen eingetroffenen Briefen hat er tatsächlich diese Gebiete bereist und weilte dort längere Zeit. Nun kehrte er wieder in seine Vaterstadt Bennisch in Schlesien zurück und diese Rücksicht gab Anlaß zu einer großen Ehrung für den Heimgekehrten, der nunmehr mit den größten Forschern in gleiche Reihe gestellt werden kann. Der junge Forscher wurde von der Stadtvertretung, den Vereinen und der Bevölkerung festlich empfangen.

relevancy

text

In der holländischen Hafenstadt Rotterdam und in ganz Holland erregt das Schicksal der russischen Rückwanderer viel Aufsehen, die an Bord des Dampfers „Voluturno“ der Uranium=Linie aus Amerika nach Europa zurückgekommen waren und von den deutschen Behörden daran gehindert wurden, durch deutsches Gebiet nach Rußland zu reisen. Diese 56 Russen konnten nämlich die von den preußischen Behörden verlangten Dokumente (Durchfahrkarten bis Wirballen und russische Grenzpässe) nicht vorweisen. Die armen Leute mußten also, als der „Voluturno“ Rotterdam verließ, wieder an Bord gehen, der Kapitän aber ließ sie nochmals ausschiffen. Jetzt werden die Rückwanderer auf Kosten der Uranium=Dampfschiffahrts=Gesellschaft verpflegt. Wie sich ihr Schicksal weiter gestalten wird, muß sich erst noch entscheiden.

3

relevancy

text

In der holländischen Hafenstadt Rotterdam und in ganz Holland erregt das Schicksal der russischen Rückwanderer viel Aufsehen, die an Bord des Dampfers „Voluturno“ der Uranium=Linie aus Amerika nach Europa zurückgekommen waren und von den deutschen Behörden daran gehindert wurden, durch deutsches Gebiet nach Rußland zu reisen. Diese 56 Russen konnten nämlich die von den preußischen Behörden verlangten Dokumente (Durchfahrkarten bis Wirballen und russische Grenzpässe) nicht vorweisen. Die armen Leute mußten also, als der „Voluturno“ Rotterdam verließ, wieder an Bord gehen, der Kapitän aber ließ sie nochmals ausschiffen. Jetzt werden die Rückwanderer auf Kosten der Uranium=Dampfschiffahrts=Gesellschaft verpflegt. Wie sich ihr Schicksal weiter gestalten wird, muß sich erst noch entscheiden.

3

In [4]:

In [5]:

```

1 %matplotlib inline
2 import matplotlib.pyplot as plt
3 df_newspaper = pd.read_csv('data/export_returnmigration_16_04_2021_21_35.csv')
4 fig =
  df_newspaper.groupby(['relevancy', 'newspaper_id']).size().unstack().plot(kind='bar',
  ,stacked=True)
5 plt.title('figure 3: Manually annotated newspapers clippings on the topic of return
  migration (0 = irrelevant, 3 = relevant).')
6 plt.show()

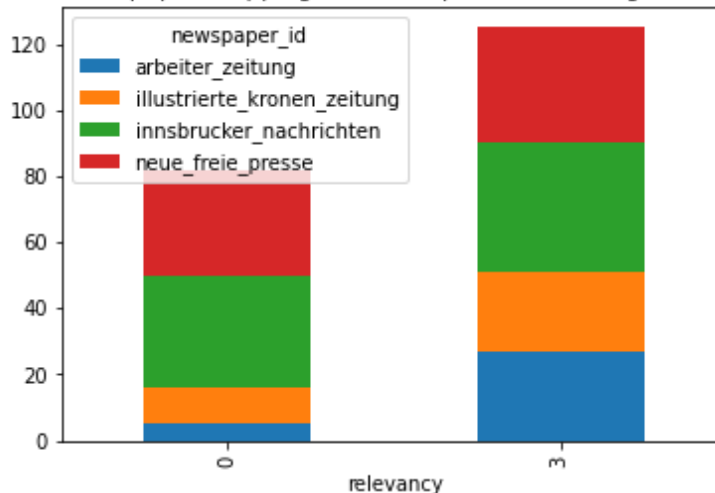
```

Outputs changed



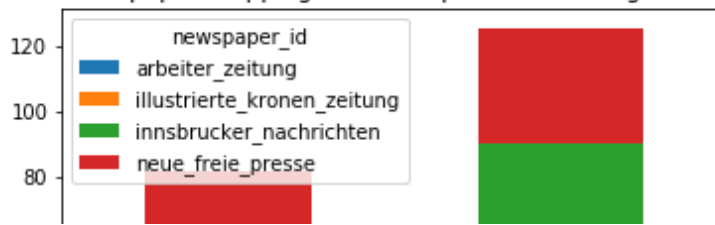
Output deleted

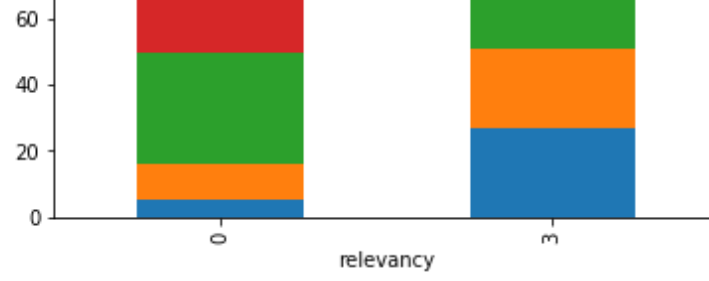
figure 3: Manually annotated newspapers clippings on the topic of return migration (0 = irrelevant, 3 = relevant).



Output added

figure 3: Manually annotated newspapers clippings on the topic of return migration (0 = irrelevant, 3 = relevant).





In [5]:

In [6]:

```

13
14 #remove stop words
15 nltk.download('stopwords')
16 nltk.download('punkt')
17
18 stop_words = stopwords.words('german')
19 #add stop words manually
20 stop_words.extend(["a",
"ab","aber","ach","acht","achte","achten","achter","achtes","ag","alle","allein","a
llem","allen","aller","allerdings","alles","allgemeinen","als","also","am","an","an
dere","anderen","ändern","anders","au","auch","auf","aus","ausser","außer","ausserd
em","außerdem","b","bald","bei","beide","beiden","beim","beispiel","bekannt","berei
ts","besonders","besser","besten","bin","bis","bisher","bist","c","d","da","dabei",
"dadurch","dafür","dagegen","daher","dahin","dahinter","damals","damit","danach","d
aneben","dank","dann","daran","darauf","daraus","darf","darfst","darin","darüber","
darum","darunter","das","dasein","daselbst","dass","daß","dasselbe","davon","davor"
,"dazu","dazwischen","dein","deine","deinem","deiner","dem","dementsprechend","demg
egenüber","demgemäss","demgemäß","demselben","demzufolge","den","denen","denn","den
selben","der","deren","derjenige","derjenigen","dermassen","dermaßen","derselbe","d
erselben","des","deshalb","desselben","dessen","deswegen","d.h","dich","die","dieje
nige","diejenigen","dies","diese","dieselbe","dieselben","diesem","diesen","dieser"
,"dieses","dir","doch","dort","drei","drin","dritte","dritten","dritter","drittes",
"du","durch","durchaus","dürfen","dürft","durfte","durften","e","eben","ebenso","eh
rlich","ei","ei","eigen","eigene","eigenen","eigener","eigenes","ein","einander","
eine","einem","einen","einer","eines","einige","einigen","einiger","einiges","einma
l","eins","elf","en","ende","endlich","entweder","er","Ernst","erst","erste","erste
n","erster","erstes","es","etwa","etwas","euch","f","früher","fünf","fünfte","fünft
en","fünfter","fünftes","für","g","gab","ganz","ganze","ganzen","ganzer","ganzes","
gar","gedurft","gegen","gegenüber","gehabt","gehen","geht","gekannt","gekonnt","gem
acht","gemocht","gemusst","genug","gerade","gern","gesagt","geschweige","gewesen","
gewollt","geworden","gibt","ging","gleich","gott","gross","groß","grosse","große",
"grossen","großen","grosser","großer","grosses","großes","gut","gute","guter","gutes
","h","habe","haben","habt","hast","hat","hatte","hätte","hatten","hätten","heisst"
,"her","heute","hier","hin","hinter","hoch","i","ich","ihm","ihn","ihnen","ihr","ih
re","ihrem","ihren","ihrer","ihres","im","immer","in","indem","infolgedessen","ins"
,"irgend","ist","j","ja","jahr","jahre","jahren","je","jede","jedem","jeden","jeder
","jedermann","jedermanns","jedoch","jemand","jemandem","jemanden","jene","jenem",
"jenen","jener","jenes","jetzt","k","kam","kann","kannst","kaum","kein","keine","kei
nem","keinen","keiner","kleine","kleinen","kleiner","kleines","kommen","kommt","kön
nen","könnt","konnte","könnte","konnten","kurz","l","lang","lange","leicht","leide"
,"lieber","los","m","machen","macht","machte","mag","magst","mahn","man","manche",
"manchem","manchen","mancher","manches","mann","mehr","mein","meine","meinem","meine
n","meiner","meines","mensch","menschen","mich","mir","mit","mittel","mochte","möch
te","mochten","mögen","möglich","mögt","morgen","muss","muß","müssen","musst","müss
t","musste","mussten","n","na","nach","nachdem","nahm","natürlich","neben","nein","
neue","neuen","neun","neunte","neunten","neunter","neuntes","nicht","nichts","nie",
"niemand","niemandem","niemanden","noch","nun","nur","o","ob","oben","oder","offen"
,"oft","ohne","Ordnung","p","q","r","recht","rechte","rechten","rechter","rechtes",
"richtig","rund","s","sa","sache","sagt","sagte","sah","satt","schlecht","Schluss",
"schon","sechs","sechste","sechsten","sechster","sechstes","sehr","sei","seid","sei
en","sein","seine","seinem","seinen","seiner","seines","seit","seitdem","selbst","s
ich","sie","sieben","siebente","siebenten","siebenter","siebentes","sind","so","sol
ang","solche","solchem","solchen","solcher","solches","soll","sollen","sollte","sol
lten","sondern","sonst","sowie","später","statt","t","tag","tage","tagen","tat","te
il","tel","tritt","trotzdem","tun","u","über","überhaupt","übrigens","uhr","um","un
d","und?","uns","unser","unsere","unserer","unter","v","vergangenen","viel","viele"
,"vielen","vielleicht","vier","vierte","vierten","vierter","viertes","vom"
,"von","vor","w","wahr?","während","währenddem","währenddessen","wann","war","wäre"
,"waren","wart","warum","was","wegen","weil","weit","weiter","weitere","weiteren",
"weiteres","welche","welchem","welchen","welcher","welches","wem","wen","wenig","wen
ige","weniger","weniges","wenigstens","wenn","wer","werde","werden","werdet","wesse
n","wie","wieder","will","willst","wir","wird","wirklich","wirst","wo","wohl","woll
en","wollt","wollte","wollten","worden","wurde","würde","wurden","würden","x","y","

```



```
[nltk_data] Downloading package stopwords
[nltk_data]   /Users/elisabeth.guerard/
[nltk_data] Package stopwords is already up-to-date
[nltk_data] Downloading package punkt to
[nltk_data]   /Users/elisabeth.guerard/
[nltk_data] Package punkt is already up-to-date
```

```
[nltk_data] Downloading package stopwords
[nltk_data]   C:\Users\c62255\AppData\Roaming\nltk_data
[nltk_data] Package stopwords is already up-to-date
[nltk_data] Downloading package punkt to
[nltk_data]   C:\Users\c62255\AppData\Roaming\nltk_data
[nltk_data] Package punkt is already up-to-date
```

table 4: Relevancy, original text and pre-processed

relevancy	text
3	<p>[Oesterreichisch=ungarische Natural=Verpflegsstation in Hamburg.] Heute tritt in Hamburg die vom Oesterreichisch=ungarischen Hilfsvereine errichtete NaturalVerpflegsstation ins Leben. In derselben befinden sich vorläufig sechs Betten, die aber bei dem starken Andränge hilfesuschender Oesterreicher und Ungarn, insbesondere der Rückwanderer aus überseeischen Ländern, beiweitem nicht ausreichend sind, die wünschenswerthe Vergrößerung der Verpflegsstation kann erst erfolgen, sobald der Verein die dazu nöthigen Geldmittel aufzubringen in der Lage sein wird. Die in der Verpflegsstation untergebrachten Landsleute und auch solche, welche dort nicht untergebracht werden können, erhalten außerdem eine einfache, aber ausreichende Nahrung (Frühstück, Mittag= und Abendbrot), wogegen die Unterstützung bedürftiger Reisender mit Bargeld von heute an eingestellt wurde.</p>

table 4: Relevancy, original text and pre-processed

relevancy	text
3	<p>[Oesterreichisch=ungarische Natural=Verpflegsstation in Hamburg.] Heute tritt in Hamburg die vom Oesterreichisch=ungarischen Hilfsvereine errichtete NaturalVerpflegsstation ins Leben. In derselben befinden sich vorläufig sechs Betten, die aber bei dem starken Andränge hilfesuschender Oesterreicher und Ungarn, insbesondere der Rückwanderer aus überseeischen Ländern, beiweitem nicht ausreichend sind, die wünschenswerthe Vergrößerung der Verpflegsstation kann erst erfolgen, sobald der Verein die dazu nöthigen Geldmittel aufzubringen in der Lage sein wird. Die in der Verpflegsstation untergebrachten Landsleute und auch solche, welche dort nicht untergebracht werden können, erhalten außerdem eine einfache, aber ausreichende Nahrung (Frühstück, Mittag= und Abendbrot), wogegen die Unterstützung bedürftiger Reisender mit Bargeld von heute an eingestellt wurde.</p>

relevancy	text
	[Die Bilanz von Monte=Carlo.] Der Mailänder Secolo veröffentlicht die Bilanz von Monte=Carlo, wie sie mit Schluß des Finanzjahres (31. October) aufgestellt wurde. Das Erträgniß der Spielbank betrug 14.850,000 Francs (im vergangenen Finanzjahre 19.850,000 Francs). Ausgaben: Civilliste für den Fürsten Albert von Monaco 2.000,000 Francs; Polizei, Gendarmerie, Unterricht und öffentliche Anlagen 1.500,000 Francs; Directoren, Verwalter, Croupiers und 0 Dienstpersonal 1.000,000 Francs; Theater, Orchester, Rennen, Regatten, Taubenschießen und Wohlthätigkeit 800,000 Francs; Annoncen und Einschaltungen 500,000 Francs; Reisegelder für verunglückte Spieler, um ihnen die Heimkehr zu ermöglichen, 100,000 Francs; ebensoviel wurde auch für die Verhinderung von Selbstmorden verausgabt. Nichtsdestoweniger haben sich im abgelaufenen Finanzjahre 35 Personen wegen ihrer Spielverluste entleibt.

relevancy	text
	[Die Bilanz von Monte=Carlo.] Der Mailänder Secolo veröffentlicht die Bilanz von Monte=Carlo, wie sie mit Schluß des Finanzjahres (31. October) aufgestellt wurde. Das Erträgniß der Spielbank betrug 14.850,000 Francs (im vergangenen Finanzjahre 19.850,000 Francs). Ausgaben: Civilliste für den Fürsten Albert von Monaco 2.000,000 Francs; Polizei, Gendarmerie, Unterricht und öffentliche Anlagen 1.500,000 Francs; Directoren, Verwalter, Croupiers und 0 Dienstpersonal 1.000,000 Francs; Theater, Orchester, Rennen, Regatten, Taubenschießen und Wohlthätigkeit 800,000 Francs; Annoncen und Einschaltungen 500,000 Francs; Reisegelder für verunglückte Spieler, um ihnen die Heimkehr zu ermöglichen, 100,000 Francs; ebensoviel wurde auch für die Verhinderung von Selbstmorden verausgabt. Nichtsdestoweniger haben sich im abgelaufenen Finanzjahre 35 Personen wegen ihrer Spielverluste entleibt.

In [6]:

In [7]:

```

1 #create testing and training corpus
2 np.random.seed(1)
3 msk = np.random.rand(len(df)) < 0.599
4 train_df = df[msk]
5 train_df.reset_index(drop=True,inplace=True)
6 test_df = df[~msk]
7 test_df.reset_index(drop=True,inplace=True)

```

In [7]:

In [8]:

```

1 #plot the result
2 my_colors = [(0.20,0.200,0.50), (0.100, 0.75, 0.200)] #set colors
3 fig, axes = plt.subplots(1,2,figsize=(10,3))
4 test_df.relevancy.value_counts().plot(kind='bar', color = (0.100, 0.75, 0.20),
5 ax=axes[1])
6 test_df.relevancy.value_counts().plot(kind='bar', color = my_colors, ax=axes[1])
7 train_df.relevancy.value_counts().plot(kind='bar', color = (0.100, 0.75, 0.20),
8 ax=axes[0])
9 train_df.relevancy.value_counts().plot(kind='bar', color = my_colors, ax=axes[0])
10 axes[1].legend(['Non_Relevant', 'Relevant'])
11 axes[0].legend(['Non_Relevant', 'Relevant'])
12 axes[1].title.set_text('figure 4a: Testing Corpus.')
13 axes[0].title.set_text('figure 4b: Training Corpus.')
14 print(f"The training corpus contains {len(train_df)} articles,
15 {train_df.relevancy.value_counts()[3]} of which are relevant and
16 {train_df.relevancy.value_counts()[0]} irrelevant.")
17 print(f"The test corpus consists of {len(test_df)} articles,
18 {test_df.relevancy.value_counts()[3]} of which are relevant and
19 {test_df.relevancy.value_counts()[0]} irrelevant.")

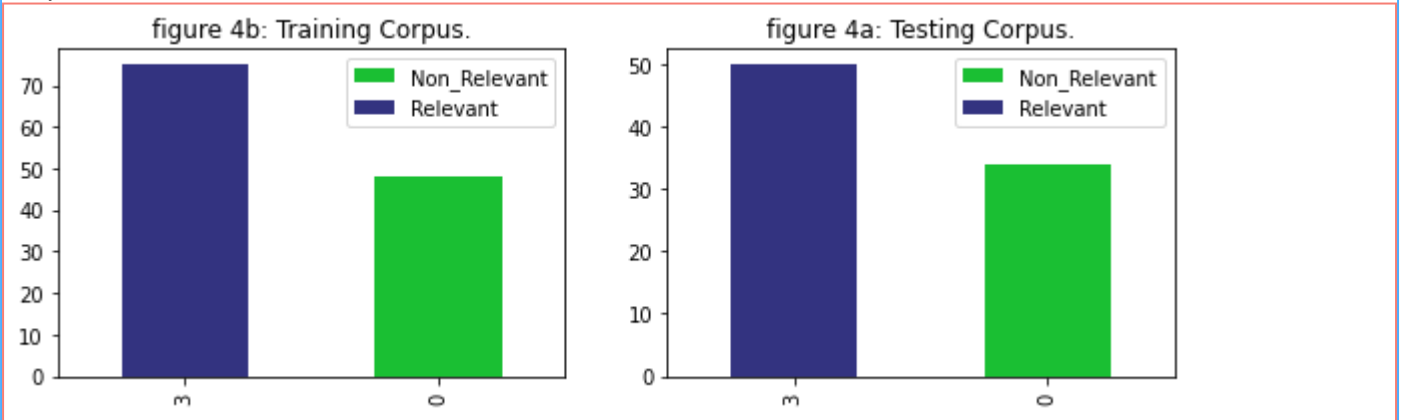
```

Outputs changed



The training corpus contains 123 articles, 75 of which are relevant and 48 of which are irrelevant.
The test corpus consists of 84 articles, 50 of which are relevant and 34 of which are irrelevant.

Output deleted



Output added

