

The initial research plan was to study two art historical journals: Finnish *Tahiti* and *South African Journal of Art Historians* (from here onwards referred as *SAJAH*), which both are significant art historical journals in their countries.¹ The plan was motivated by the knowledge that art historical research has been historically applied for constructing national narratives in both countries by the ruling elite, which has been challenged in the past decades for differing reasons. As a result art historical discipline has fragmented nationally and globally, but some national trends in art historical writing can still be observed. The original plan was to research the most common topics discussed in each journal and compare similarities and differences of the topics discovered to study the state of art history today in the historically diverging countries.

However, the preliminarily gathering data from the issues published between 2011 and the first issue of 2017 revealed that *Tahiti* and *SAJAH* are hardly comparable: *Tahiti* consist research articles, columns, book and exhibition reviews to name a few, whereas *SAJAH* consists predominantly research articles.² Furthermore, out of the 256 papers published in *Tahiti* c. 82% are in Finnish, 11% in Swedish, 6% in English and one in Danish. In comparison 223 papers published in *SAJAH* during the same period 97,3% are published in English and 2,7% in Afrikaans.³ All gathered dataset are also included into the repository. Due to the variation of languages published in *Tahiti* as well the obvious problem of directly comparing the journals, the initial research plan was reformed. All gathered dataset are also included into the repository.

As the initial study showed papers published in *SAJAH* form a more uniform corpus in terms of language used and types of articles. Therefore, the research focuses solely on *SAJAH*. The primary research questions are: what are the current topics of South African art historical research? Has there been a change in research topics in post-apartheid South Africa when demands for decolonising education and academic research has been increasingly made? Decolonising art history, noted by many scholars, is a significant issue in South Africa, as South African art history has been historically based on Western models and has primarily focused on Western art and a concept of art resulting in neglecting so called African art and local art done by racialized groups.⁴ Thus, the aim of this research is to study what are the major topics discussed in South African art history and whether South African art history is "localised". The study is conducted by applying topic modelling (Gensim LDA) on a corpus of articles published in *SAJAH* between 2011 and 2017 to extract topics discussed in papers. Topic modelling is applied as a try-out to test whether a humanities scholar trained in close reading can apply digital methods to conduct a research. Furthermore, reading all the 223 articles, or their abstracts and titles, to grasp their subject matter would be time-consuming, thus topic modelling is tested in this project to study whether it can speed the research process.

¹ *Tahiti* is published by Finnish Society of Art Historians and *South African Journal of Art Historians* by Art Historical Work Group of South Africa. Editions from 2011 onwards of both *Tahiti* [<https://tahiti.journal.fi/>] and *South African Journal of Art historian* [<https://sajah.co.za/>] can be found on Internet.

² Only an obituary and a few book reviews were observed during the gathering of data.

³ Issues studied were *Tahiti*: 2011 vol1 no 1 — 2017 vol 7 no 1, *SAJAH*: 2011 vol 26 no 2 — 2017 vol 32 no 1. In *Tahiti* there are 256 articles, 210 of them in Finnish, 29 in Swedish, 16 in English and one in Danish. Together articles consist 583 528 words: in Finnish 437 000 (74,9%), in Swedish 84 187 (14,4%) and in English 62 341 (10,7%). Note Danish was not included in word count. In *SAJAH* there are 223 articles, 217 in English and 6 in Afrikaans. Together articles consist 1 089 901 words: in English 1 049 662 (96,3%) and in Afrikaans 40 239 (3,8%).

⁴ See e.g. Becker's dissertation of the possibility of decolonising the art historical discipline in various South African art institutions (2017); Carman (2007) on historical exclusion of black artist and art historian from the discipline; Freschi (2012) on the historical base of South African art history in Western model and discussion of the changes in post-apartheid era; Bach et al. (2006) on critical discussion of South African art history in a global context.

Materials and methods

The research material consist of all the articles published in English in *SAJAH* between the second issue 2011 (vol 26 no 2) and the first issue of 2017 (vol 32 no 1). The first issue of 2011 was not available due to linking error on the website of *SAJAH*, which is the reason why the first issue of 2017 was included but not the rest to get issues from "full six years", consisting 19 issues. The number of issues published has varied from two to four per year, similar to articles published: for example in 2012 issue 2 consist of 21 articles whereas 2013 issue 1 consist of only one longer research article. Further irregularity is that some issues are dedicated to a particular topics like 'situated experience' (2015, no 3) and 'encounter in visual arts' (2016 no 1) directing research articles published. Also six articles published in Afrikaans out of 223 published articles were left out of the research. All the issues are available on the website of *SAJAH* in pdfs (<https://sajah.co.za/>). The articles were manually copy-pasted to edit out titles, subheadings, abstracts, captions, notes and bibliography. In other words, only the body text was extracted to the corpus to decrease the possibility of biased results caused for example by inclusion of bibliographies or abstracts. The corpus of articles was saved in plain text format forming a dataset which was further preprocessed in Python (attachment: *sajah_english.txt*).

The further preprocessing of the dataset and analysis were conducted by applying two different codes. The process was conducted with an assistance of a civil engineer C. Nordberg. The two topic modelling codes applied were based on published codes on Internet, slightly modified by us. Applying two different codes was conducted as we both were novices in computational text analysis and could not fully judge the validity of the codes applied or their results. Thus, applying two separate codes to the dataset was to limit potential biases and limitations of one. However, further studying and evaluation of the codes is necessarily needed. The first code is modification of Susan Li's code (https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/topic_modeling_Gensim.ipynb) and the second on Radim Rehůřek's, the creator of Gensim, (https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html). In this article modified version of Lin's code is titled *python1* and Rehůřek's *python2* (attachments: *python1.py* and *python2.py*).

Python1

With *python1* the dataset is first cleaned and tokenized.

```
import spacy
spacy.load('en')
from spacy.lang.en import English
parser = English()
def tokenize(text):
    lda_tokens = []
    tokens = parser(text)
    for token in tokens:
        if token.orth_.isspace():
            continue
        elif token.like_url:
            lda_tokens.append('URL')
        elif token.orth_.startswith('@'):
            lda_tokens.append('SCREEN_NAME')
        else:
            lda_tokens.append(token.lower_)
    return lda_tokens
```

And then lemmatized to get the basic form of words.

```
import nltk
nltk.download('wordnet')
from nltk.corpus import wordnet as wn
def get_lemma(word):
    lemma = wn.morphy(word)
    if lemma is None:
        return word
    else:
        return lemma

from nltk.stem.wordnet import WordNetLemmatizer
```

```
def get_lemma2(word):
    return WordNetLemmatizer().lemmatize(word)
```

Further the stopwords list from nltk is applied to the dataset to eliminate the most common used English words.

```
nltk.download('stopwords')
en_stop = set(nltk.corpus.stopwords.words('english'))

def prepare_text_for_lda(text):
    tokens = tokenize(text)
    tokens = [token for token in tokens if len(token) > 3]
    tokens = [token for token in tokens if token not in en_stop]
    tokens = [get_lemma(token) for token in tokens]
    return tokens
```

Also the words less than 3 digits are filtered out, and the data is prepared for Latent Dirichlet Allocation (LDA) by transforming data to a vectorized form.

```
from gensim import corpora
dictionary = corpora.Dictionary(text_data)
corpus = [dictionary.doc2bow(text) for text in text_data]
import pickle
pickle.dump(corpus, open('corpus.pkl', 'wb'))
dictionary.save('dictionary.gensim')
```

The LDA model was executed 10 times, asked to identify 20 topics, do 100 passes and print 20 words of each topic. Different variations of passes, number of topics and words printed were tested, and after several try-outs the above mentioned were concluded to be satisfactory: 100 passes took 700—1000s, and 20 topics and words give a general view of the dataset, albeit very limited. To compare the results of each code, the same numbers were also applied to code python2.⁵

```
NUM_TOPICS = 20
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics = NUM_TOPICS, id2word=dictionary,
passes=100)
# ldamodel.save(output_location + '/' + filename + '.gensim')
ldamodel.save(file_name + '.gensim')
topics = ldamodel.print_topics(num_words=20)
return topics
```

Visualisation with the code was not conducted in this research due to limited time.

Python2

With the python2 the dataset is cleaned in more detail: together with tokenizing, the text is converted in lower cases, numeric tokens and tokens consisting only one character are filtered out. Similarly to python1, words are lemmatized to get their basic form.

```
from nltk.tokenize import RegexpTokenizer

tokenizer = RegexpTokenizer(r'\w+')
for idx in range(len(docs)):
    docs[idx] = docs[idx].lower()
    docs[idx] = tokenizer.tokenize(docs[idx])

docs = [[token for token in doc if not token.isnumeric()] for doc in docs]

docs = [[token for token in doc if len(token) > 1] for doc in docs]

from nltk.stem.wordnet import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()
docs = [[lemmatizer.lemmatize(token) for token in doc] for doc in docs]
```

⁵ See detailed explanation of the model in Susan Li's 'Topic Modelling in Python with NLTK and Gensim' (2018) on Towards Data Science website [<https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21>] (last accessed 3.1.2020).

Bigrams are added in order to keep significant "word pairs" together. In the dataset used for example 'South Africa' is a bigram that repeats constantly in the dataset and reading them separately could potential bias the findings.

```
from gensim.models import Phrases

# Add bigrams and trigrams to docs (only ones that appear 20 times or more).
bigram = Phrases(docs, min_count=20)
for idx in range(len(docs)):
    for token in bigram[docs[idx]]:
        if ' ' in token:
            # Token is a bigram, add to document.
            docs[idx].append(token)
```

Further, based on their frequency the most common and rare words are removed. An additional stopwords list was added as the original code yielded results with very general words, like 'it', 'with', 'for', hindering the interpretation of topics, potentially suggesting that the code should be reorganised.

```
from gensim.corpora import Dictionary
dictionary = Dictionary(docs)
dictionary.filter_extremes(no_below=20, no_above=0.5)

en_stop = set(nltk.corpus.stopwords.words('english'))
docs = [[token for token in doc if token not in en_stop] for doc in docs]
```

Similarly to python1, text is transformed in vectorized form for LDA modelling, and the same numbers to python1 code are applied for topics, number of words and passes. Iteration is 400 following the original code but chunksize is determined as 11000 as the number of documents is 10646 to speed the process (in comparison number of unique tokens is 4569). Usefully the code counts topic coherence of each topic by using "Umass" topic coherence measure". The initial try-outs of assigning 20 topics yielded on average -3— -1 topic coherence, 30, 40 lower negative results similarly 10, 12 and 15, finally resulting in a decision to set the number of topics in 20 (in the result shown in this study the average topic coherence is -2.8578).

```
corpus = [dictionary.doc2bow(doc) for doc in docs]

from gensim.models import LdaModel

num_topics = 20
chunksize = 11000
passes = 100
iterations = 400
eval_every = None

temp = dictionary[0] # This is only to "load" the dictionary.
id2word = dictionary.id2token

model = LdaModel(
    corpus=corpus,
    id2word=id2word,
    chunksize=chunksize,
    alpha='auto',
    eta='auto',
    iterations=iterations,
    num_topics=num_topics,
    passes=passes,
    eval_every=eval_every
)

top_topics = model.top_topics(corpus) #, num_words=20)

avg_topic_coherence = sum([t[1] for t in top_topics]) / num_topics
print('Average topic coherence: %.4f.' % avg_topic_coherence)
```

```
from pprint import pprint
pprint(top_topics)

print('Number of unique tokens: %d' % len(dictionary))
print('Number of documents: %d' % len(corpus))
```

Findings

The complete lists of topics generated can be found in attached files titled python1findings.xlsx and python2findings.xlsx containing 20 topics each with 20 separate words. I will briefly introduce here the findings with both codes and then discuss on the findings. Note that random words from the topic wordlists are included in the introduction.

Python1

A number of topics identified in analysis contain only generic words that could be part of any art historical research (topics 2, 7, 12, 17) like words 'time, experience, world, present, meaning...' in topic 2. Topic 11 includes words like 'wife, female, child, home, privacy, mother, family, jung, archetype' which suggest handling potentially role of females and males combined with Jungian archetypes. Discussion on gender and gender roles is relatively common globally in art history due to the strong influence of feminist theory, and its is common in art history, again globally, to integrate various theories from different disciplines in art historical research, which could explain inclusion of Jung's archetypes. Furthermore, discussion on gender is relatively common in South Africa across disciplines due to the traditionally patriarchal society which is growingly constested. However, from the listed words it is difficult to tell their connection to art history i.e. in which context they appear, whether it is South African art or society or a global discussion on gender.

Similarly, several topics can be identified to concern architecture but their further identification is difficult with given words (topics 3, 9, 14, 19, word examples of topic 19: 'house, wall, building, figure, roof, stone, form'). Some topics on architecture can be identified slightly further: topic 6 includes words like 'architecture, ruin, greek, building, classical, renaissance, architect, century' suggesting concentration on European ancient and classical architecture and topic 15 containing words like 'pleasure, plant, garden, temporality, schulz, pallasmaa' which points to planning that involves acknowledging the role of nature which is crucially important in works of internationally renowned architects Juhani Pallasmaa and Christian Norberg-Schulz. Also topic 5, including words like 'painting, depict, symbol, colour, painter, represent, body, bosch', can be identified a little further: it could be understood concerning painting, Bosch potentially referring to the 15th century Dutch painter Hieronymus Bosch.

Out of the 20 topics, nine topics contain words that refer certainly to the local context. Topic 1 ('first, south, town, africa, years, become, cape, know, early, time, century, ndebele) might deal with the Ndebele community which was historically strong kingdom in Southern Africa prior the British colonisation, and whose rich cultural traditions have been studied for decades in South African art history. Topic 18 contains also reference to a South African ethnic group Tswana ('traditional, kgotla, tswana, area, city, urban, space, public, market'), potentially dealing with integration of traditional ways of living in urban spaces, also much discussed topic in South African art history and social sciences (kgotla is traditional Botswanan meeting place in the village, also traditional to certain groups in South Africa). Topics 8 and 13 hint towards architecture: topic 8 (monument, memorial, voortrekker, afrikaner, decrease, blood, "moerdyk, burgher) to Gerard Moerdijk's now much contested Voortrekker Monument and topic 13 (corbusier, fagan, concrete, vernacular) to modern architecture and potentially Gabriel Fagan, one of the most acknowledged South African modern architects. However, whether the topics are handled critically cannot be judged from the list of words given. Topic 4 ('identity, cultural, south, african, political, history, society, colonial, culture, memory') could include critical examination of identity politics and politics of memory which are much discussed topics also in South African art history on the post-apartheid era. Similarly, topic 16 ('indian, boer, british, afrikaner, race, religion, racial, english, gender, judgement, anglo, adorno') indicated discussion on identity politics, specially race and ethnicity. Topic 20 ('statue, peace, reconciliation, president, leaders, kruger, motive, weapon') might also refer similar post-apartheid discussion. Finally, topics 10 ('leonardo, campaign, copy, derrida, venice, biennale, heidegger, violence, 1993, animal, king, admit, abuse, William, maintenance...') contain such a variety of words that its exact identification is difficult.

Python2

Similarly to the python1, analysis with python2 offer five topics (topics 1, 3, 5, 7, 9, 18) that contains such common words that they could form any kind of art historical research, like words 'time, past, present, experience, one' in topic 1 and words 'body, space, figure, human, landscape, one, form, painting'. Similar to topic 11 in python1, topic 14 contains words ('women, men, ruin, space, body, home, social, ritual, life, public, resistance, female, child') referring to discussion of gender and potentially gender roles at home and in society, yet missing words referring to Jung's theory. Again, several topics contain words suggesting topics dealing with architecture with rather general terms (topics 10, 12, 15, e.g. words in topic: 'house, wall, figure, building, roof', 'floor'), topic 10 referring to one of the most famous modern architects Le Corbusier who is frequently mentioned whenever modern architecture is discussed. Also topic 4 contains words ('architect, architecture, le, corbusier, design, building, regionalism, local, university, critical') suggesting a topic dealing with Le Corbusier, potentially in South African context as local interpretations of modern architecture are much discussed theme in South African architectural history, often termed regionalism (earlier mentioned Fagan being the pioneer in it). Also comparable to topic 6 in python1, topic 6 includes words ('building, temple, site, structure, street, design, greek, church') that hint towards European (ancient) architecture.

Further similarities to findings of python1 are topics 16 and 17. Topic 16 has words like 'environment, landscape, community, area, plant, garden, people, urban, kgotla, cultural', similar to topic 18 in python1, whereas topic 17 ('house, material, stone, ndebele, site, homestead, design, settlement, dwelling, pattern, construction, wall, village') resembles topic 1 in python1, yet missing words referring to the past. Another topic referring to a native cultural group in South Africa is found in python2: topic 19 contains a word xhosa together with words 'experience, et al, state, one, object, derrida, consciousness, brain, flow' — xhosa could be refer to an ethnic group or the language, however the group of words given make it hard to interpret what is the topic about. Topic 11 differs also from the findings of python1 containing words like 'art, artist, work, painting, van, der, merwe, painter', potentially concerning a local painter. South African is also present in topics 8 and 13: topic 8 ('south, african, africa, town, century, south_africa, cape, city, colonial, white, british, people, european, cape_town, industrial') potentially dealing with the colonial history of Cape Town and topic 13 ('war, first, year, became, family, boer, would, british, van, camp, river...') with the conflicts between Afrikaaners and British in the late 19th century.

Topic 2 contains words like 'monument, memory, history, memorial, apartheid, political, historical, museum, past, statue, site, africa, south_africa, voortrekker, afrikaner, post' which might suggest a topic dealing with politics of memory and history, Apartheid Museum in Johannesburg and potentially already mentioned Voortrekker monument and its problems in post-apartheid South Africa. Unlike topic 8 in python1 in which Voortrekker monument was also mentioned, this topics indicates that the presence of the monument dedicated to Afrikaaners in the post-apartheid society might be critically discussed in the journal. Finally, topic 20 contains also 'venice and biennale' like topic 10 in python1, but also 'image, photograph, work, water, film, kentridge, photograph' most likely referring to a internationally known South African artist William Kentridge who has exhibited in Venice Biennale several times, also in 1993.

Discussion

The application of two separate codes and their findings have certain similarities in words and topics. In both of the codes roughly half of the findings contain topics with very general words which makes their identification difficult. Architecture is strongly present in found topics and around half of the topics could be interpret to deal with South African subjects suggesting that South African art history is to certain extend localised. Within the few names included in the wordlists some (white) South African artist and architects are mentioned like Kentridge, Fagan and Moerdijk along international famous scholars and artist like Derrida, Heidegger and Le Corbusier. Furthermore, some words like Ndebele, Xhosa and Tswana suggest that South African ethnic groups other than white are included in the studies. However, they appear mostly in connections with words describing ways of living which can hardly be considered revolutionary — the 20th century anthropology studied in detail lifestyle of the above mentioned groups and also South African art historians have been making inventories of their material culture already in the apartheid era. In short, the findings suggest that in SAJAH local art historical topics are discussed

along topics concerning Western art, but there is no clear indication of fully decolonising the discipline, that is to bring African intellectuals and artist to the core of the discipline.

However, in both results of the codes a few topics indicate that there might be some critical discussion on identity politics and politics of memory and history. Yet, topic 8 in python1 and topic 2 in python2 both contain words voortrekker and monument but the other words differ which demonstrates how significant the other terms appearing in the list are for interpreting the topic. Topic 8 (python1) could be interpreted to discuss solely on the history of Voortrekker monument whereas words in topic 2 (python2) indicate towards more critical discussion of the monument. The differing findings demonstrate that before making any final conclusion of the state of South African art history, a further study of LDA model of Gensim should be conducted, not only to optimise the topic coherence measure, but also to broaden the understanding of how topic model works and how to optimise its accuracy. Also, further study of each library, like spaCy, applied should be also conducted as the preprocessing the dataset is crucial and as shown here, gives a differing results which can lead into biased interpretation.

Moreover, in the initial phase of manually gathering the data some human errors might have occurred in copy-pasting leading potentially into duplicates, unwanted or missing data. Even though I attempt to correct word which were divided into syllables some might have remained in the dataset biasing the findings. Further, converting the dataset into different formats with Adobe Acrobat, might have caused distortion of the text data which should be examined. In preprocessing the text data in Python, a greater care should be paid to irregular white spaces and for example terms 'figure' and 'image' which are often applied in the body text to refer to illustrative images, usually in form (image n) or (figure n). In the findings of both codes the term 'figure' appeared frequently which might be due to the above mentioned fact. In python2 unintelligible words 'ha' and 'wa' appeared in many wordlist which suggest the code should be modified. Also some problem with detecting biagrams could be noted with python2 from the wordlists as for example 'cape, town and cape_town' all appeared in the same list. Similarly, the stopwordlist should be examined and potentially supplement or deduct some words. For the further study, and to realize the initial research plan, a closer study of preprocessing and implementation of LDA model of Gensim is needed. One of the codes applied here could be potentially modified to be suitable for the purpose if careful and knowledgeable improvement is done.

Furthermore, the size of the dataset used poses a certain limitation to draw any final conclusions of the state of South African art history today. For a future research more recent issues of SAJAH should be included and potentially expand the timescale to pre-2011 issues to get more comprehensive view. Also the articles in Afrikaans excluded from this study should be considered to be added for analysis. Additionally, even though SAJAH, established in 1983, is a significant peer reviewed art historical journal in South Africa, other sources should be included for further research if a holistic view of the state of South African art history is aimed to achieve. A brief survey of SAJAH's issues from the 1980's and the 90's shows how some the same authors from apartheid era are still contributing to the journal.⁶ Similarly, the preliminary familiarisation with the issues used in this study shows that majority of the articles in different issues are predominantly written by the same authors whose research interests directs the findings of this study. For example, exceptionally many, in comparison to *Tahiti*, contributors of the journal focus on studying architecture which can be seen on the results of this research. Alike individual researcher, like Estelle Alma Maré whose research concentrates on mainly European antique or renaissance art, has contributed nearly every issue which impacts on the findings, similarly to the issues dedicated to a specific themes. Therefore, a further study of Art Historical Work Group of South Africa publishing SAJAH and its representativeness of South African art historians is needed before concluding anything about the state of South African art history. According to my shallow knowledge of South African art history, there is an increasing number of South African scholars who are challenging the existing view of art history but whose work is not published in SAJAH.⁷ Including their studies could yield results which would show more inclusion of local and "African"

⁶ See past issue of SAJAH for example in the repository of University of Pretoria [<https://repository.up.ac.za/handle/2263/10058>] (last accessed 4.1.2020).

⁷ The assumption of not including the contradictory authors is based on reading the names of the published authors which are predominantly English or Afrikaans names representing the minority of the population.

art and decolonisation of the discipline. Thus, this study is mere a study of SAJAH and its representation of South African art history in the 2010's.

The research suggest that digital methods offer a potentially useful way of getting insight into one's dataset specially when using a large corpus. However, deeper knowledge and skills of digital methods is certainly needed before applying digital methods as part of one's research or drawing any final conclusion of the dataset. As art historian Murtha Baca et al. note in their introduction to digital art history in *Visual Resources* (2019): "several of the papers make very clear, technology is not a "magic bullet," and working with digital tools and methods is not "easier" or faster than traditional art history (though of course many computations can be done at speeds that would be unthinkable using manual methods). Data sets (including images) in all of their messy, incomplete, and sometimes seemingly contradictory nature must be well understood, and carefully "curated" and prepared for analysis in order to be able to draw any kind of valid conclusions."⁸

Literature

Baca, Murtha, Anne Helmreich & Melissa Gill, 2019. Digital Art History, *Visual Resources*, 35:1-2, 1-5 (DOI: 10.1080/01973762.2019.1556887).

Bach, Fredrich Teja, James Elkins, Andrea Giunta, Ladislav Kesner, Sandra Kloppe & David Summer, 2006. The Art Seminal. *Is Art History Global?* Ed. James Elkins. New York: Routledge, 113—176.

Becker, Danielle Loraine, 2017. *South African Art History: A possibility of decolonising a discourse*. Cape Town: University of Cape Town.

Carman, Jillian, 2007. South Africa: Empowering the Local. *Global and Local Art Histories*. Ed. Gregory Minissale & Celina Jeffery. Newcastle upon Tyne: Cambridge Scholars Publishing, 59—83.

Freschi, Federico, 2012. Other Views: Art History in (South) Africa and the Global South. *Diogenes*, 58(3) 93–101 (DOI: 10.1177/039219211245646).

⁸ Baca et al. 2019, 3.