

What I've learnt about Open Science by being a part of CEVOpen

Perspectives on Open Science

Shweata N. Hegde,
Project Manager
(CEVOpen) and
NIPGR Intern

Presentation structure

1. What we @ CEVOpen do?



2. What is Open Science?



3. My message and learnings!

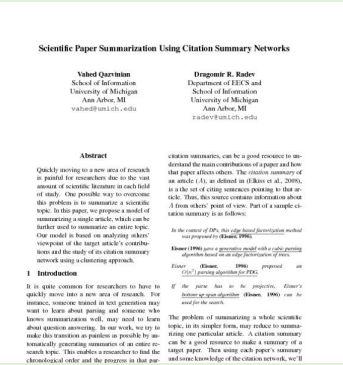


What we @CEVOpen do



Gain Insights

1000s plants
more!



Unstructured

PMCID	Chemicals	plants

Structured



Lantana Camara (An Invasive)



Tulsi

What we @CEVOpen do



Make
interpretations

Scientific Paper Summarization Using Citation Summary Networks

Yahel Quesada
School of Information
University of Michigan
Ann Arbor, MI
yquesa@umich.edu

Dragomir R. Radev
Department of EECS and
School of Information
University of Michigan
Ann Arbor, MI
radev@umich.edu

Abstract
Quickly moving to a new area of research is painful for researchers due to the vast amount of scientific literature in each field of study. One possible way to overcome this problem is to summarize a scientific topic. In this paper, we propose a model of summarizing a single article, which can be further used to summarize an entire topic. Our model is based on analyzing what viewpoint of the target article's contribution and the study of its citation summary network using a summarization approach.

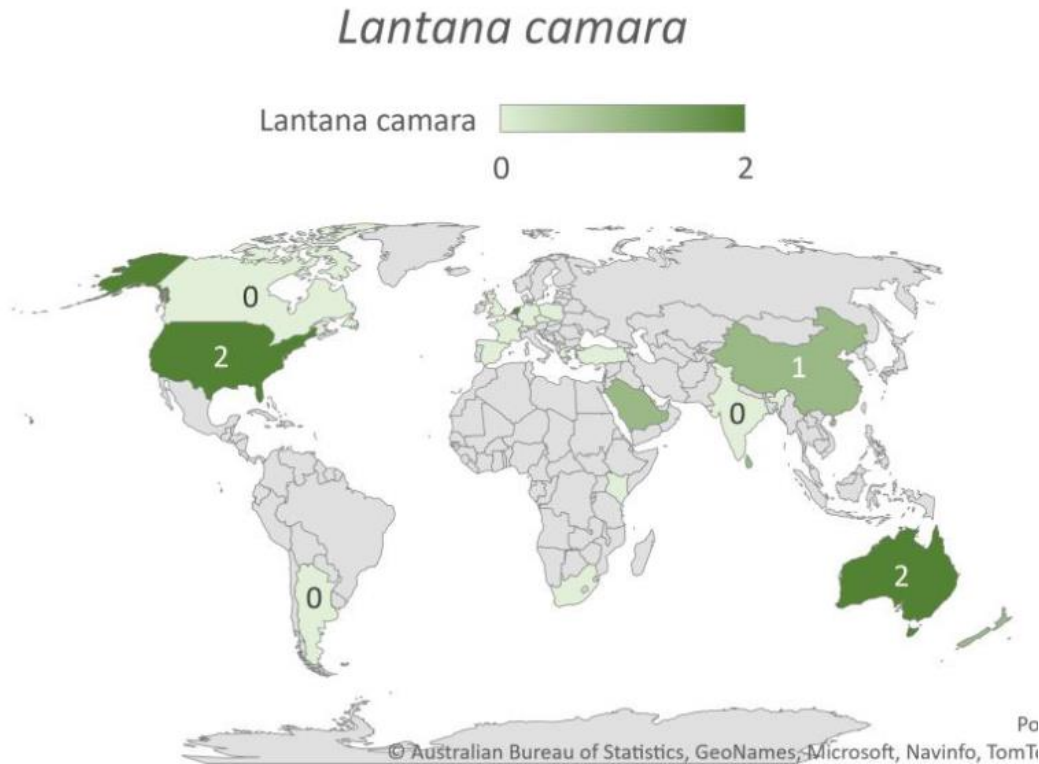
1 Introduction
It is quite common for researchers to have to quickly move into a new area of research. For instance, someone trained in text generation may want to learn about mining and someone who knows summarization well, may need to learn about question answering. In our work, we try to make this transition as painless as possible by automatically generating summaries of an entire research topic. This enables a researcher to find the chronological order and the progress in that

research area, can be a good resource to understand the main contribution of a paper and how the paper affects others. The citation summary of an article (1), as defined in (Radev et al., 2006), is a list of citing sentences pointing to that article. Thus, this source contains information about a from where" point of view. Part of a sample citation summary is as follows:
In the context of (2), this paper-based summarization method was proposed by (Radev, 2006).
Radev (2006) uses a generative model with a topic-pointing algorithm based on an edge distribution of nodes.
Radev (2006) proposed an "edge" pointing algorithm for (2006).
If the paper has to be presented, Radev's summarization algorithm (Radev, 2006) can be used for the search.
The problem of summarizing a whole scientific topic, in its simplest form, may reduce to summarizing one particular article. A citation summary can be a good resource to make a summary of a target paper. Thus using each paper's summary and some knowledge of the citation network, we'll

location	Chemicals	plants

Unstructured

Structured



representation

What we @CEVOpen do



Reduce time it takes to review literature

Months to review literature!



Rapidly download bulk of papers



DONE
Minutes, if not hours!

Analyse and annotate



Scientific Paper Summarization Using Citation Summary Networks

Vahed Qazvinian
School of Information
University of Michigan
Ann Arbor, MI
vahed@umich.edu

Dragomir R. Radev
Department of EECS and
School of Information
University of Michigan
Ann Arbor, MI
radev@umich.edu

Abstract

Quickly moving to a new area of research is painful for researchers due to the vast amount of scientific literature. One possible way to overcome this problem is to summarize a scientific topic. In this paper, we propose a model of summarizing a single article, which can be further used to summarize an entire topic. Our model is based on analyzing others' viewpoint of the target article's contributions and the study of citation summary network using a clustering approach.

Citation summaries, can be a good resource to understand the main contributions of a paper and how that paper affects others. The *citation summary* of an article (A), as defined in (Elkiss et al., 2008), is a set of citing sentences pointing to that article. Thus, this source contains information about A from others. A sample citation summary is as follows:

In the context of DP, this edge based factorization method was proposed by (Khan, 1996).

Elkiss (1996) gives a generative model with a cubic parsing algorithm based on an edge factorization of area.

Elkiss (Khan, 1996) proposed an $O(n^3)$ parsing algorithm for PEG.

If the parse has to be projective, Elmer's bottom-up parse algorithm (Khan, 1996) can be used for the search.

1 Introduction

It is quite common for researchers to have to quickly move into a new area of research. For instance, a new text generation may want to learn about parsing, and someone who knows summarization well, may need to learn about question answering. In our work, we try to make this transition as painless as possible by automatically generating summaries of an entire research topic. This enables a researcher to find the chronological order and the progress in that par-

The problem of summarizing a whole scientific topic, in its singular form, may reduce to summarizing one particular article. A citation summary can be a good response to make a summary of a target paper's summary and some knowledge of the citation network, we'll

Visions for CEVOpen

Accessibility!
Non-native English
speakers



Ask ami!



What's
Ocimum

Language	Ocimum tenuiflorum (Q960124)
Kannada	ತುಳಸಿ
Malayalam	തൂളസി
Marathi	तुळस
Odia	ତୁଳସୀ ଗଛ
Punjabi	ਤੁਲਸੀ
Sanskrit	तुलसी
Tamil	துளசி
Telugu	తులసి
Hindi	तुलसी
Assamese	তুলসী
Gujarati	તુલસી
Urdu	تلسی

From Wikidata

Ocimum species represent commercially important medicinal and aromatic plants. The essential oil biosynthesized by *Ocimum* species ...

Visions for CEVOpen

De-jargonizing
science



Phenyl-
propanoids?

Look at ami's
annotation!

Ocimum species represent commercially important medicinal and aromatic plants. The essential oil biosynthesized by Ocimum species is enriched with specialized metabolites specifically, terpenoids and phenylpropanoids.

Phenylpropanoid

From Wikipedia, the free encyclopedia

The **phenylpropanoids** are a diverse family of organic compounds that are synthesized by plants from the amino acids phenylalanine and tyrosine.^[1] Their name is derived from the six-carbon, aromatic phenyl group and the three-carbon propene tail of coumaric acid, which is the central intermediate in phenylpropanoid biosynthesis. From 4-coumaroyl-CoA emanates the biosynthesis of myriad natural products including lignols (precursors to lignin and lignocellulose), flavonoids, isoflavonoids, coumarins, aurones, stilbenes, catechin, and phenylpropanoids.^[2] The coumaroyl component is produced from cinnamic acid.

WE make use of, and also contribute to...



Open literature 30% of all literature*



Open database - Wikidata



Open source tools

Impossible if ALL literature was closed and copyrighted

*reference

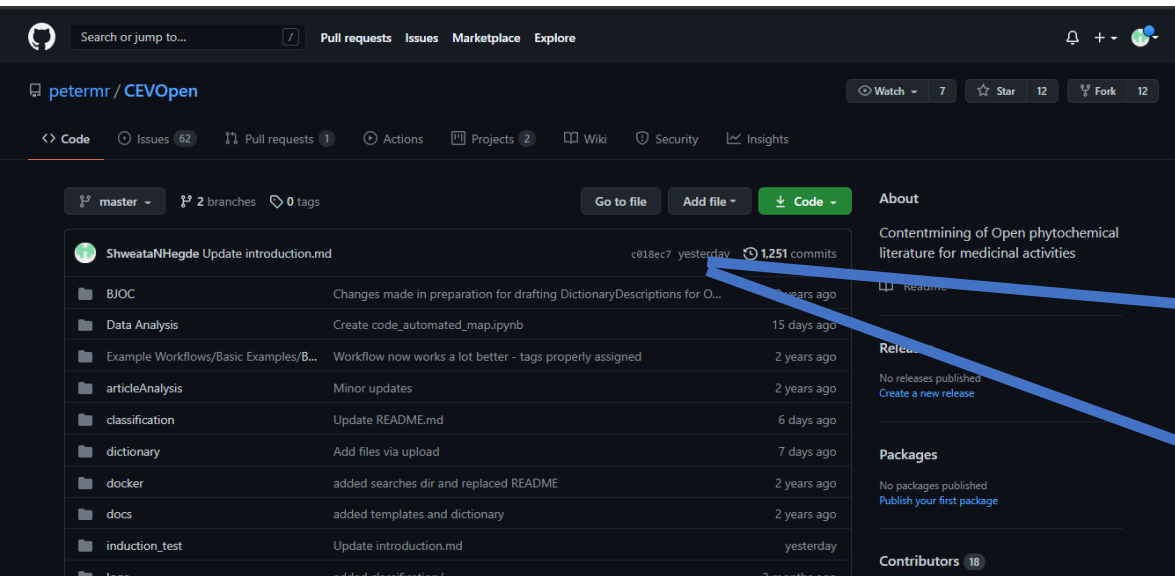
Open Science is
more than *just* open
publishing...

It's also about...



1.

How we work



updated in real-time! GITHUB

Try it!

<https://github.com/petermr/CEVOpen>

<https://github.com/petermr/pygetpapers>

<https://github.com/petermr/pyami>

Open Notebook Philosophy*

*https://en.wikipedia.org/wiki/Open-notebook_science

Community and Accountability



Not-started

In-progress

Completed

ISSUES

ISSUES	Ambreen	Vaishali	Priya	Rajan	Vanisha	Sana	Kareena	Charles
corpus (between 200-950)	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Ready
dictionary	Ready	Ready	Ready	Ready	Ready	Ready	Ready	In progress
annotated (Pos/Neg) viral epidemics	8.90%	40/10	In progress	111/116	In progress	36/14	121/29	5/35
specialist subsets (e.g. methods) e.g. sections	Ready	Ready	Ready	Ready	Ready	Ready	Ready	Haven't Started
notebooks	Ready	In progress	In progress	In progress	In progress	Haven't Started	In progress	Haven't Started
machine learning / NLP	In progress	In progress	Haven't Started	In progress	In progress	In progress	In progress	In progress
display	In progress	Haven't Started	Haven't Started	Haven't Started	Haven't Started	Haven't Started	In progress	In progress
language variants	Ready	Haven't Started	Ready	Ready	Haven't Started	In progress	Haven't Started	Haven't Started
ami search	Ready	In progress	Ready	Ready	Ready	Ready	Ready	Haven't Started



2.

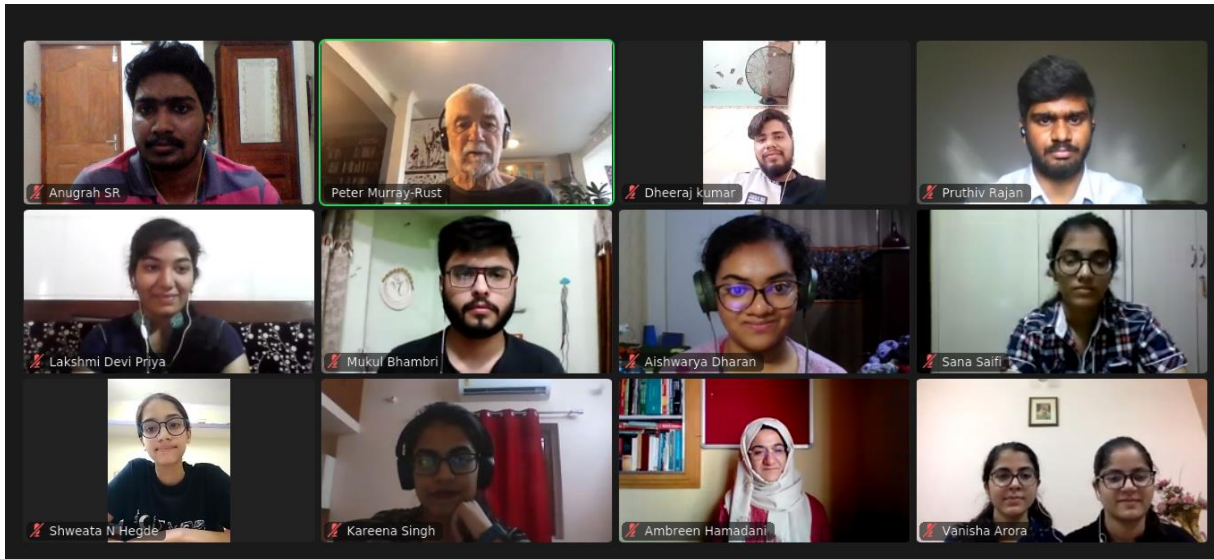
Work as a community
Individual accountability

It's also about...



3.

Whom we include and collaborate with



cambiohack

Outreach!
serendipitous encounters!

The Team! Scattered across the country



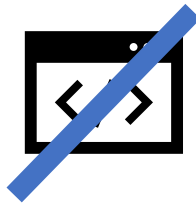
Transition to Open Science
practice in *ANY* ways
possible or **promote!**

Context



Life Science Undergrad

Joined openVirus -> 11 months ago (Sep. 2020)



No coding experience!

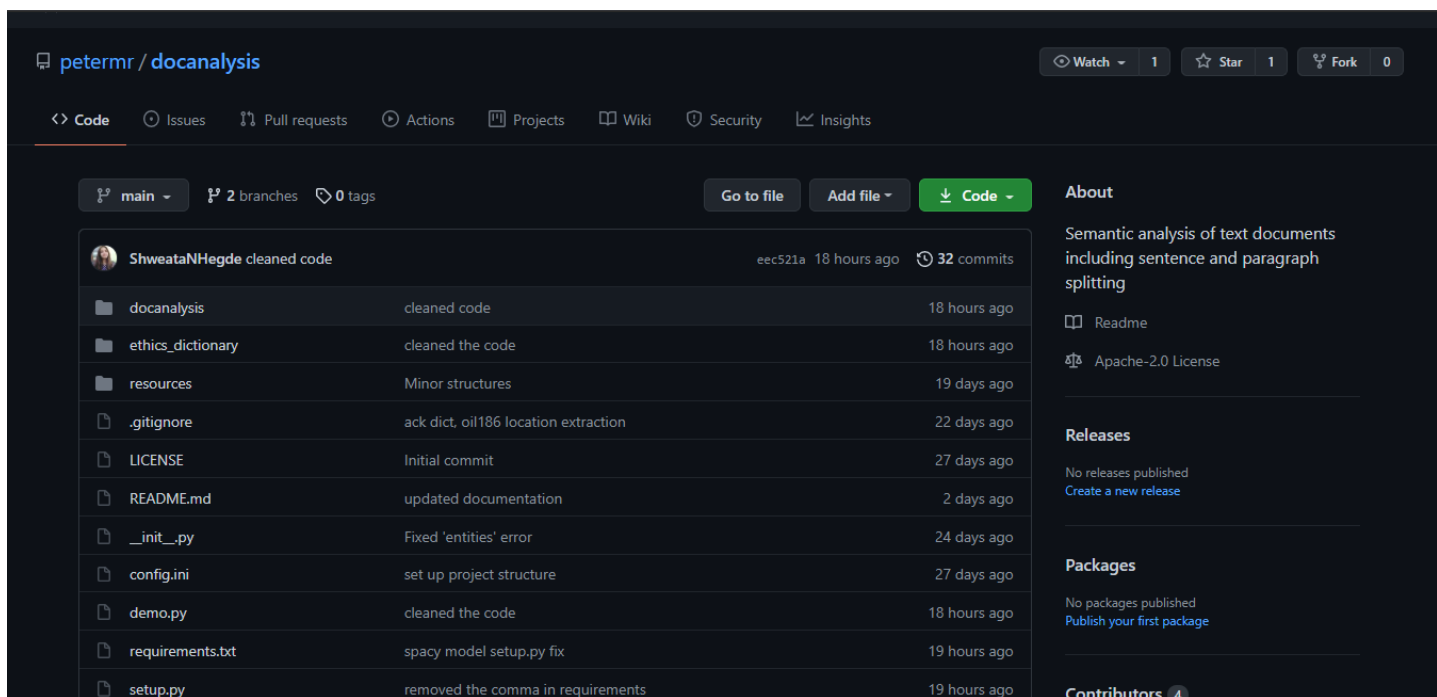


You don't have to be
an expert to start out!

No restrictions!
No gatekeeping!

Present Project:

Unsupervised entity extraction from scientific literature



The screenshot shows the GitHub repository 'petermr/docanalysis'. The repository has 1 watch, 1 star, and 0 forks. The main branch is selected, showing 2 branches and 0 tags. The file list includes:

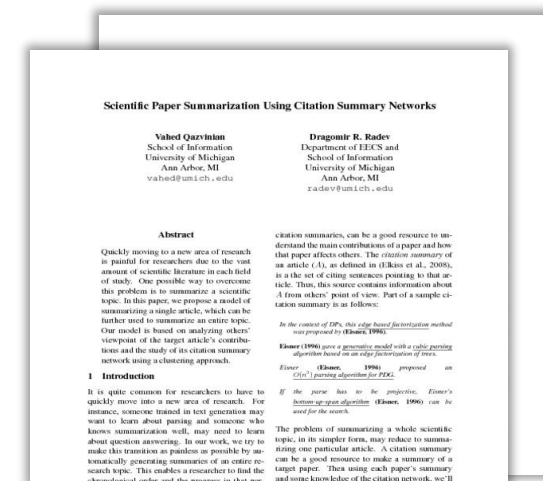
File/Folder	Description	Time
docanalysis	cleaned code	18 hours ago
ethics_dictionary	cleaned the code	18 hours ago
resources	Minor structures	19 days ago
.gitignore	ack dict, oil186 location extraction	22 days ago
LICENSE	Initial commit	27 days ago
README.md	updated documentation	2 days ago
init.py	Fixed 'entities' error	24 days ago
config.ini	set up project structure	27 days ago
demo.py	cleaned the code	18 hours ago
requirements.txt	spacy model setup.py fix	19 hours ago
setup.py	removed the comma in requirements	19 hours ago

The repository also includes an 'About' section with a semantic analysis of text documents, a 'Readme', and an 'Apache-2.0 License'. There are no releases or packages published yet.

Docanalysis

<https://github.com/petermr/docanalysis>

Working with Ayush Garg



PMCID	Ethics Committee

Reflecting on learnings!

Documenting your work **openly** helped me...

 Forgo *perfectionism*: Everything's work-in-progress

 Experiment: Document failures!

Reflecting on learnings!

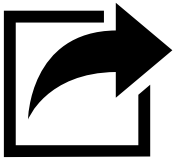
The **open science community** has helped me...



Learn: Coding, best-practices from others!



Repurpose: Don't reinvent!



Share: Without worries!



Communicate better: Reading and writing documentation



Long way to go!
Grateful that OpenVirus and CEVOpen are a part
of my journey!

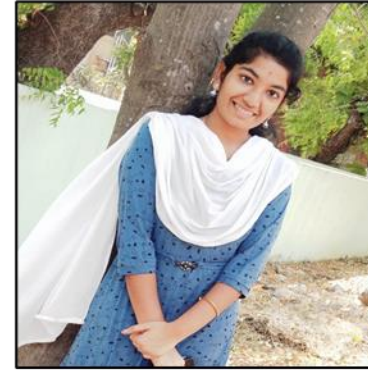
Our Team



Dr Peter Murray-Rust
*Reader Emeritus,
University of Cambridge*



Dr Gitanjali Yadav
*NIPGR (India), University
of Cambridge (UK)*



Lakshmi Devi Priya
*BTech Student, Govt
College of Technology*



Kareena Singh
*MSc Scholar,
Fergusson College*



Dr. Ambreen Hamadani
*PhD Scholar, Veterinarian
SKUAST-K*



Sana Saifi
*B.Sc. Student
University of Delhi*

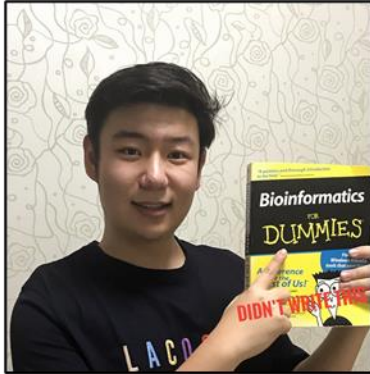


Pruthivrajan Karunakaran
*BTech Student
Dr. MGR Edu. & Res. Institute*



Vaishali Arora
*MSc Scholar,
University of Delhi*

Our Team



Charles Zeyang Li
*BA Hons, University of
Cambridge*



Vanisha Arora
*MSc Scholar,
Osmania University*



Dheeraj Dhingani
*BSc Student
RR College Alwar*



Pooja Pareek
*MSc Scholar, Maharaja
Ganga Singh University*



Urja Biswas
*MSc Scholar, University
of Rajasthan*



Om Prakash Mehra
*MSc Scholar, MSJ college
Bharatpur, Rajasthan*



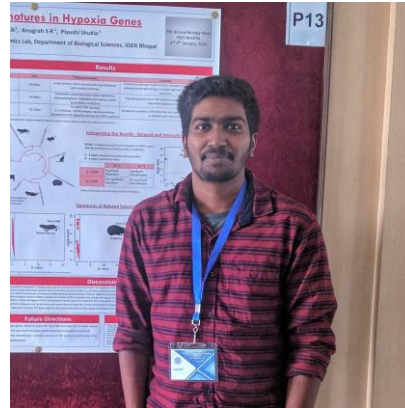
Jitu Ram Bhargav
*MSc Scholar,
University of Rajasthan*



Simranleen Singh
*MSc (Chemistry),
Fergusson College*



Aishwarya Dharan
MSc (Bioinformatics)
Central Univ. Punjab



Anugrah S. R.
BS-MS(Biological
Sciences) IISER Bhopal



Shweeta N. Hegde
BSc Student
Regional Inst. Of Edu.
Mysore



Ayush Garg
GIIS Singapore



Mukul Bhambri
Undergrad, SRM
University



Vasant Kumar
M.Sc. Biotech,
Himachal Pradesh
University, Summerhill



Talha Hasan
MSc Toxicology
Jamia Hamdard
University



Radhu Kantilal Ladani
Msc Bioinformatics
Rajiv Gandhi
Institute, Pune.



Chaintanya Sharma
B.Tech(Mechanical)
Delhi Technological University



Kanishka Parashar
MSc Biotech
Jamia Millia Islamia, Delhi



Sagar Jadhav
Postdoctoral Research
Associate, NIPGR

Many other global
volunteers!

Lead on to...

Sagar's work on TPS

Chaitanya's work on Acknowledgements