A Clustering-Based Algorithm for Automatic Document Separation

Project ID: 29

1. GITHUB Link

https://github.com/TheIndianCoder/A-Clustering-Based-Algorithm-for-Automatic-Document-Separation

2. Team Members

★ Mehtaab Singh (20171034)
★ Amrit Preet Singh(2019909001)

3. Project Goals

- To construct a model for estimating inter-page similarity in ordered collections of document images, based on a combination of text and layout features.
- Implement a discriminative classifier and segment the document images based on the chosen decision rule .

4. Problem Definition

Input : A sequence of disjoint groups of images.

Processing : To use features based on layout document structure and topic concepts to discriminate between related and unrelated images.

Output : Clusters of related images into documents.

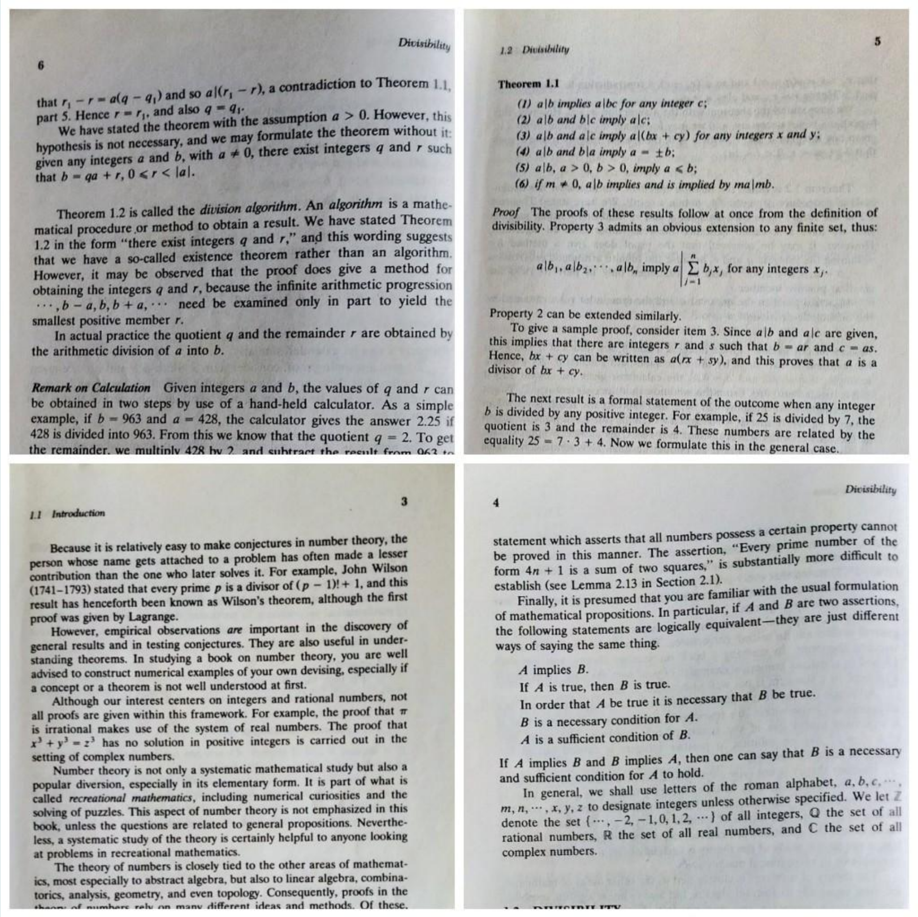
Terminology : The term document signifies an ordered collection of images . A single image in a document is termed as a page . Two pages are “related” if they come from the same underlying document and are “unrelated” otherwise.

5. Results of the project

- Our model should match the expectation of the paper being referred (“Thompson.et.al 2002”) .

- In other words , the model should succeed at successfully clustering similar pages together into a document and separate the unrelated pages with high accuracy.

Example of expected outcome:



that $r_1 - r = a(q - q_1)$ and so $a|(r_1 - r)$, a contradiction to Theorem 1.1, part 5. Hence $r = r_1$, and also $q = q_1$.

We have stated the theorem with the assumption $a > 0$. However, this hypothesis is not necessary, and we may formulate the theorem without it: given any integers a and b, with $a \neq 0$, there exist integers q and r such that $b = qa + r$, $0 \leq r < |a|$.

Theorem 1.2 is called the *division algorithm*. An *algorithm* is a mathematical procedure or method to obtain a result. We have stated Theorem 1.2 in the form "there exist integers q and r," and this wording suggests that we have a so-called existence theorem rather than an algorithm. However, it may be observed that the proof does give a method for obtaining the integers q and r, because the infinite arithmetic progression $\cdots, b-a, b, b+a, \cdots$ need be examined only in part to yield the smallest positive member r.

In actual practice the quotient q and the remainder r are obtained by the arithmetic division of a into b.

Remark on Calculation Given integers a and b, the values of q and r can be obtained in two steps by use of a hand-held calculator. As a simple example, if $b = 963$ and $a = 428$, the calculator gives the answer 2.25 if 428 is divided into 963. From this we know that the quotient $q = 2$. To get the remainder, we multiply 428 by 2, and subtract the result from 963 to

Theorem 1.1

(1) $a|b$ implies $a|bc$ for any integer c;
(2) $a|b$ and $b|c$ imply $a|c$;
(3) $a|b$ and $a|c$ imply $a|(bx + cy)$ for any integers x and y;
(4) $a|b$ and $b|a$ imply $a = \pm b$;
(5) $a|b$, $a > 0$, $b > 0$, imply $a \leq b$;
(6) if $m \neq 0$, $a|b$ implies and is implied by $ma|mb$.

Proof The proofs of these results follow at once from the definition of divisibility. Property 3 admits an obvious extension to any finite set, thus:

$$a|b_1, a|b_2, \cdots, a|b_n \text{ imply } a \,\Big|\, \sum_{j=1}^{n} b_j x_j \text{ for any integers } x_j.$$

Property 2 can be extended similarly.

To give a sample proof, consider item 3. Since $a|b$ and $a|c$ are given, this implies that there are integers r and s such that $b = ar$ and $c = as$. Hence, $bx + cy$ can be written as $a(rx + sy)$, and this proves that a is a divisor of $bx + cy$.

The next result is a formal statement of the outcome when any integer b is divided by any positive integer. For example, if 25 is divided by 7, the quotient is 3 and the remainder is 4. These numbers are related by the equality $25 = 7 \cdot 3 + 4$. Now we formulate this in the general case.

Because it is relatively easy to make conjectures in number theory, the person whose name gets attached to a problem has often made a lesser contribution than the one who later solves it. For example, John Wilson (1741–1793) stated that every prime p is a divisor of $(p-1)! + 1$, and this result has henceforth been known as Wilson's theorem, although the first proof was given by Lagrange.

However, empirical observations *are* important in the discovery of general results and in testing conjectures. They are also useful in understanding theorems. In studying a book on number theory, you are well advised to construct numerical examples of your own devising, especially if a concept or a theorem is not well understood at first.

Although our interest centers on integers and rational numbers, not all proofs are given within this framework. For example, the proof that π is irrational makes use of the system of real numbers. The proof that $x^3 + y^3 = z^3$ has no solution in positive integers is carried out in the setting of complex numbers.

Number theory is not only a systematic mathematical study but also a popular diversion, especially in its elementary form. It is part of what is called *recreational mathematics*, including numerical curiosities and the solving of puzzles. This aspect of number theory is not emphasized in this book, unless the questions are related to general propositions. Nevertheless, a systematic study of the theory is certainly helpful to anyone looking at problems in recreational mathematics.

The theory of numbers is closely tied to the other areas of mathematics, most especially to abstract algebra, but also to linear algebra, combinatorics, analysis, geometry, and even topology. Consequently, proofs in the theory of numbers rely on many different ideas and methods. Of these,

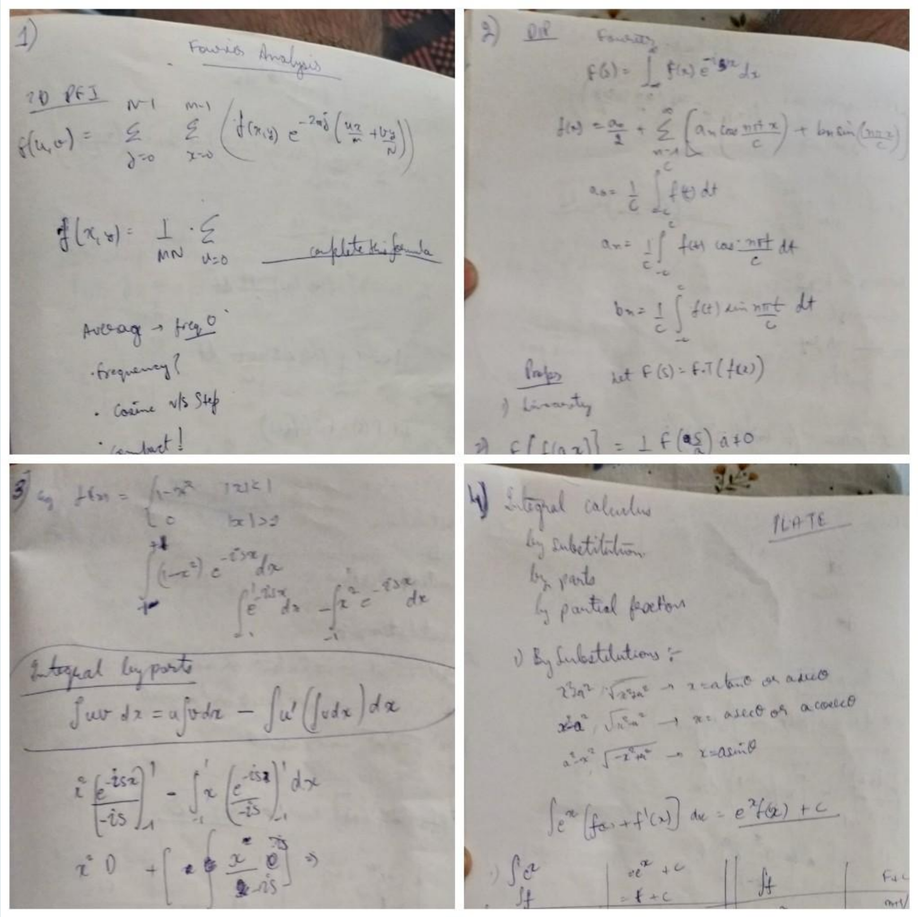
statement which asserts that all numbers possess a certain property cannot be proved in this manner. The assertion, "Every prime number of the form $4n + 1$ is a sum of two squares," is substantially more difficult to establish (see Lemma 2.13 in Section 2.1).

Finally, it is presumed that you are familiar with the usual formulation of mathematical propositions. In particular, if A and B are two assertions, the following statements are logically equivalent—they are just different ways of saying the same thing.

A implies B.
If A is true, then B is true.
In order that A be true it is necessary that B be true.
B is a necessary condition for A.
A is a sufficient condition of B.

If A implies B and B implies A, then one can say that B is a necessary and sufficient condition for A to hold.

In general, we shall use letters of the roman alphabet, $a, b, c, \cdots, m, n, \cdots, x, y, z$ to designate integers unless otherwise specified. We let $\mathbb{Z}$ denote the set $\{\cdots, -2, -1, 0, 1, 2, \cdots\}$ of all integers, $\mathbb{Q}$ the set of all rational numbers, $\mathbb{R}$ the set of all real numbers, and $\mathbb{C}$ the set of all complex numbers.



1) Fourier Analysis

2D DFT

$$f(u,v) = \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} \left(f(x,y)\, e^{-2\pi j \left(\frac{ux}{M} + \frac{vy}{N} \right)} \right)$$

$$f(x,y) = \frac{1}{MN} \cdot \sum_{u=0}$$ complete the formula

Average → freq 0
· frequency?
· Cosine v/s Step
· contrast!

2) DIP Fourier

$$F(s) = \int f(x)\, e^{-j2\pi s x}\, dx$$

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{c} \right) + b_n \sin\left(\frac{n\pi x}{c} \right)$$

$$a_0 = \frac{1}{c} \int_{-c}^{c} f(t)\, dt$$

$$a_n = \frac{1}{c} \int_{-c}^{c} f(t) \cos \frac{n\pi t}{c}\, dt$$

$$b_n = \frac{1}{c} \int_{-c}^{c} f(t) \sin \frac{n\pi t}{c}\, dt$$

Props let $F(s) = F.T(f(x))$
1) Linearity
2) $F[f(ax)] = \frac{1}{a} F\left(\frac{s}{a}\right)$ $a \neq 0$

3) eg $f(x) = \begin{cases} 1 - x^2 & |x| < 1 \\ 0 & |x| > 1 \end{cases}$

$$\int_{-1}^{1} (1 - x^2)\, e^{-isx}\, dx$$

$$\int_{-1}^{1} e^{-isx}\, dx - \int_{-1}^{1} x^2 e^{-isx}\, dx$$

Integral by parts

$$\int uv\, dx = u \int v\, dx - \int u' \left(\int v\, dx \right) dx$$

4) Integral calculus PLATE

by substitution
by parts
by partial fraction

i) By Substitutions :-

$x^2 + a^2$, $\sqrt{x^2 + a^2}$ → $x = a\tan\theta$ or $a\sec\theta$
$x^2 - a^2$, $\sqrt{x^2 - a^2}$ → $x = a\sec\theta$ or $a\,\mathrm{cosec}\,\theta$
$a^2 - x^2$, $\sqrt{a^2 - x^2}$ → $x = a\sin\theta$

$$\int e^x \left[f(x) + f'(x) \right] dx = e^x f(x) + c$$

Timelines:-

1. WEEK 1 - Data cleaning and gathering
2. WEEK 2 - Extracting layout structure features from the data set
3. WEEK 3 - Extracting text similarity features from the data set
4. WEEK 4 - Implementing the classifier for final similarity score output
5. WEEK 5 - Implementing the classifier and using the decision rule to check performance of our model.
6. WEEK 6 - Final integration of all features.

Challenges :-

1. Distinguishing between table heading and page headers.
2. Some of the pages have alternating header and footers whereas some dont mention the page number
3. Section headings are hard to distinguish from page numbers.

References :-

1. Collins-Thompson, Kevyn & Nickolov, Radoslav. (2002). A Clustering-Based Algorithm for Automatic Document Separation.
2. [Doer97] D. Doermann, H. Li and O. Kia. The detection of duplicates in document image databases. In confidence.Proceedings of the International Conference on Document Analysis and Recognition, pp. 314-318, 1997.